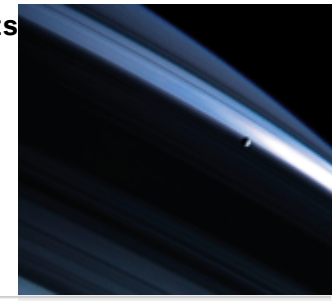
TEAM tAO

Science

Table of Contents



25 February 2005
Volume 307
Number 5713

NEW THIS WEEK:

SPECIAL ISSUE

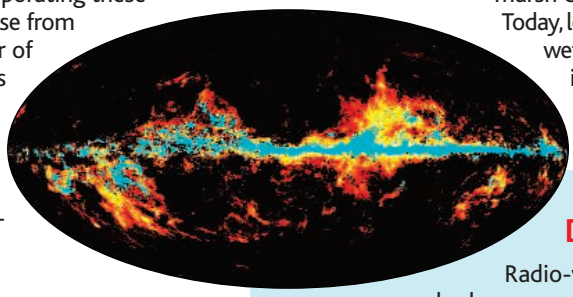
Viewpoint

Research Article

Report

New Views of Old Mars

Mars Pathfinder heralded a new generation of exploration of the red planet. Several orbiters and two rovers have successfully followed this lander, and along with telescope observations, returned information on Mars rocks and surface features, topography and gravity— which help infer its internal structure—magnetic field, and atmospheric dynamics and chemistry. **Solomon *et al.*** (p. 1214) provide an updated review incorporating these latest results and those from an increasing number of martian meteorites into a history of early Mars, including its formation and the differentiation of its core, mantle, and crust.



A Time for Planets

Most of the 130 known extrasolar planets were detected by measuring perturbations to the stellar radial velocity caused by the orbiting planet. Some planets were detected by observing decreases of the stellar light when the planet passed in front of its star. **Holman and Murray** (p. 1288) show that the timing of the transiting planet will vary if there is another planet in the system, so that the presence and mass of a second planet can be estimated from the gravitational interactions between the planets. This method may be able to detect even an Earth-mass planet.

Coupled Qubits Measured Simultaneously

For a scalable quantum computer to be realized, what will be needed is the ability to read out the entire system of qubits simultaneously, and with high fidelity. However, measurement crosstalk, an undesirable effect where the state measurement of one qubit influences that of others, has presented an experimental barrier to achieving that goal. **McDermott *et al.*** (p. 1299, see the Perspective by **Mooij**) present results on the simultaneous measurement of states of two coupled superconductor phase qubits. For the right timing sequence of the measurement, the influence of crosstalk can be minimized, and the two entangled qubits can be measured with high fidelity. The authors argue that the scheme should be applicable to multi-qubit systems.

Great Tomes of the Past Preserved

Texts often survived from Antiquity through the Middle Ages by the skin of their teeth, subject to hazards ranging from fire and war to decay and neglect. What were the odds that a manuscript would survive, that an entire work would go extinct, or that the transmission of knowledge itself could be seriously jeopardized? By treating manuscripts as though they were fossils from an extinct population, **Cisne** (p. 1305; see the Perspective by **Gilman and Glaze**) shows that explicit, testable estimates of manuscripts' and texts' survival indeed can be found under certain circumstances, and that certain works have had much greater chances of survival than has been guessed from anecdotal evidence. This work suggests a new way of using centuries' worth of exacting scholarship to investigate the survival and dissemination of information.

Restoring the Marshlands of Iraq

The marshes of southern Iraq were once the largest wetland in the Middle East and home to an indigenous population of tens of thousands of marsh dwellers. They were also a major flyway for migrating birds. Today, less than 10% of the marshes in Iraq remain as fully functioning wetlands because of extensive drainage and upstream agricultural irrigation programs on the Tigris and Euphrates rivers implemented during Saddam Hussein's regime. **Richardson *et al.*** (p. 1307) provide an assessment of the ecological status of the Iraqi marshes since the 2003 war. Nearly 20% of the original 15,000-square-kilometer marsh area was reflooded by March 2004. Reflooding has partly restored some of the former marsh areas. However, high salinity and toxicity may persist in reflooded marshes unless flow-through of fresh water is maintained by careful hydraulic design. It seems that the marshes can be restored as long as sound ecological restoration principles are followed.

Dark Gas in the Milky Way

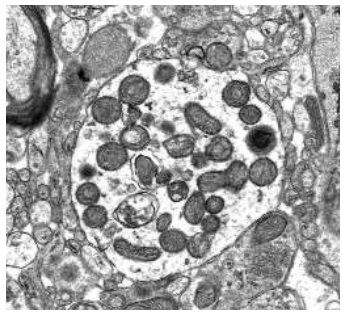
Radio-wavelength observations of atomic hydrogen and carbon monoxide (CO) have been used to estimate the abundance and distribution of gas and dust in the Milky Way Galaxy. Some cold molecular gas (dark gas) is hard to trace with radio observations, but cosmic-ray interactions with the gas can make the gas glow sufficiently to be traced. **Grenier *et al.*** (p. 1292) combined the radio and gamma-ray observations to show that the cold dark gas forms halos around CO clouds and connects these clouds to diffuse structures dominated by atomic hydrogen.

Great Medieval Earthquake

Trenching along the Main Frontal Thrust fault of the Himalayan mountains has revealed evidence for one great earthquake with a moment magnitude of about 8.8 and an offset of about 17 meters over a length of about 240 kilometers. **Lavé *et al.*** (p. 1302) dated the offset in the trench at about 1100 A.D., and they suggest that no additional great earthquakes have occurred since then. Such a great event could account for 25 to 50% of the shortening across the mountains and would recur in 1800 to 3000 years.

βAPPtists and Tauists Unite

In Alzheimer's disease (AD) pathological lesions of various types are seen in the brains of patients and in mouse models of the disease as the disease progresses. **Stokin *et al.*** (p. 1282) provide evidence that axonal transport deficits are likely to be an early characteristic of AD, contributing to progression of disease phenotypes, and perhaps being an early initiating event. The data also provide a useful unifying theme bringing amyloid precursor protein processing and tauopathy, two



CONTINUED ON PAGE 1167

CREDITS: (TOP TO BOTTOM) GRENIER *ET AL.*; STOKIN *ET AL.*

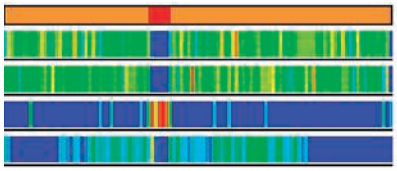
Published by AAAS

processes thought to be mechanistically distinct in neurological disease progression, together into a single disease pathway for AD.

Variety and Fitness in Natural Bacterial Populations

Microbiologists have long used clonal isolates as model systems to study processes promoted by bacterial populations. However, little is known about how functionally representative such clones may be of populations in their native habitats. **Thompson *et al.*** (p. 1311) show that vast genotypic diversity exists within a natural population of coexisting bacteria bearing nearly identical 16*S* ribosomal RNA genes (the biomarker most frequently used to identify bacterial strains). Individual clones are present in the environment at such low concentrations that, for all practical purposes, no two are alike. These findings suggest that any individual clone is relatively unimportant in overall population function and that much of the variation in the genomes does not lead to fitness differences among the members of the population.

The Bad and the Ugly?



The fungus *Cryptococcus neoformans* is an opportunistic human pathogen that has become more prevalent, in part because of the increased incidence of immunocompromised patients. **Loftus *et al***. (p. 1321, published online 13 January 2005) have sequenced the genome of two inbred strains, JEC21 and B-3501A, which differ in virulence, and have compared this sequence information with other fungal genomes. The *C. neoformans* genome shows evidence of alternative splicing and antisense transcripts, suggesting widespread genetic regulatory mechanisms, and is transposon-rich. The functional distribution of many of *C. neoformans* genes mirrors that of *Saccharomyces cerevisiae*.

Sword and Shield in Bacterial Pathogenesis

The serotype diversity that characterizes *Shigella* (and other bacterial pathogens of mammals) could have evolved under the selective pressure of the innate immune response of the host or as an escape mechanism to the adaptive response. **West *et al.*** (p. 1313; see the Perspective by **Normark *et al.***) now show that *Shigella* has acquired mechanisms to shorten the O-antigen of lipopolysaccharide (LPS) it expresses at the cell surface without decreasing its mass for the purpose of better exposing its major weapon, the type III secretory system. It does this without affecting the capacity of LPS to resist the innate immune responses of the host. Bacteriophage-mediated glucosylation of the LPS O-antigen leads to a shift from a linear to helical conformation, shortening the LPS without altering its quantity. This process provides a strong selective advantage to *Shigella* for maintaining lysogenic bacteriophages in its genome.

Mistic and Membranes

Structure determination of membrane proteins remains a challenge because of difficulties in expressing sufficient quantities of protein and in obtaining ordered crystals for analysis by x-ray crystallography. **Roosild *et al.*** (p. 1317) have taken steps forward in two directions. First, they developed techniques that allowed them to determine the structure of a four-helix bundle integral membrane protein from *Bacillus subtilis* by nuclear magnetic resonance spectroscopy. Second, they show that this protein, which they name Mistic, can be used to assist in recombinant expression of other membrane proteins because it can insert autonomously into the cell membrane.

Neuroligin and Inhibitory Synapse Formation

Achieving an appropriate balance between excitatory and inhibitory synapses as the central nervous system is wired up during development is critical for building smoothly flowing information pathways. Neuroligins are postsynaptic adhesion molecules that also play a role in synapse formation. **Chih *et al.*** (p. 1324, published online 27 January 2005; see the Perspective by **Hussain and Sheng**) have analyzed knockdown mutants of several neuroligin isoforms and found that various neuroligins have overlapping but not identical functions. Disruption of neuroligin function leads to a loss of excitatory synapses and results in a functional imbalance of excitatory and inhibitory transmission in the rodent hippocampus.

CREDIT: LOFTUS *ET AL.*

Published by AAAS

French Public Research—Saved?

On 7 January 2004, an open letter to the French government and the launch of the petition "SAUVONS LA RECHERCHE" (Save Research) started the most powerful and spontaneous protest of scientists throughout France since the 1960s. The flash point was reached after a succession of catastrophic research budgets for 3 years in a row, exacerbated by last-minute funding freezes and the elimination of 30% of the entry-level permanent research positions. This amounted to sacrificing a whole generation of young scientists. The petition gathered 75,000 signatures among working scientists and support from more than 80% of the general public. After a series of street demonstrations in French university towns, nearly half of all French laboratory directors from the main research agencies gathered on 27 March in Paris to post their resignations. A week later, after an opposition landslide in the regional elections, President Jacques Chirac disowned the research policies of his previous government and asked the newly appointed ministers for education and research to reestablish the lost academic positions, and even added 1050 university lecturer positions. After 6 months of self-organized debates in all major scientific centres, 1000 French delegates met in Grenoble on 29 October to finalize a voluminous report meant to inspire future reforms. Top government officials and national leaders of all major political parties attended this meeting.

CREDIT: LAURENT REBOURS/AP PHOTO

Undoubtedly, the longest-lasting benefit of this year-long movement was the realization by scientists and politicians alike that scientific research enjoyed unexpected strong support from the general public. "Scientific research" has been brought back into the political vocabulary and is now an electoral issue. In November, the government announced an overall 2005 budget for civilian R&D of 9.27 billion euros, a 10% increase. The number of permanent research positions in the national agencies is also maintained, and 200 temporary positions have been created to help encourage the return of foreign-based French postdocs. Yet a closer reading of this budget casts a number of shadows: One-third of the money is for fiscal measures to promote industrial R&D, and another one-third is dedicated to an ill-defined National Agency for Research. Thus, France is still only devoting a mere 0.60% of its gross national product to civilian public research, which is short of the 1% goal that a European Union directive commits us to reach by 2010 and is below the level reached in 2001 (0.74%).

There is also frustration that a year-long movement did not result in bolder proposals to reform the French academic research system and put it more in line with the organization prevailing in other leading scientific countries, including our closest European partners. In particular, the civil servant status uniquely enjoyed by French researchers from the very beginning of their careers remains unchallenged, even if it limits the number of research positions offered to postdocs and Ph.D. students and corresponds to salaries 30% lower than those offered in Germany or Switzerland. Also unchallenged is the absence of a stringent selection process for entering French universities (all of them government-funded), creating a population of rather unmotivated students. Professors are overwhelmed by teaching loads and mentoring responsibilities incompatible with serious research. Thus, despite an increase in government support, the French public research system might not retain international competitiveness without addressing these politically touchy issues. Offering more attractive (better salary, less teaching) and more numerous (but perhaps less secure) entry-level jobs will be key to retaining our most promising young scientists as well as attracting foreign-based talent.

Fortunately, things are not yet settled. The Save Research movement is gaining momentum again, after the recent release of the government's first draft of the 2006–2010 Research and Innovation Framework Act. Some scientists are concerned that the new National Agency for Research is not making enough room for curiosity-driven basic research and is keeping too much money away from existing agencies, including CNRS and INSERM. Having learned its lesson last year, the government has already postponed the presentation of the bill until June and has promised additional rounds of consultation. French research needs to be rescued, but it will require restructuring of an academic system and a government agenda by saviors on both sides.

Jean-Michel Claverie

Jean-Michel Claverie is professor at the Université de la Méditerranée School of Medicine and head of the Structural and Genomics Information Laboratory–CNRS, Marseilles, France. E-mail: Jean-Michel.Claverie@igs.cnrs-mrs.fr

10.1126/science.1109775

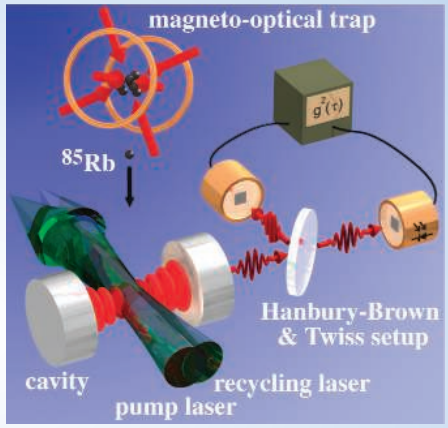
Published by AAAS

EDITORS' CHOICE

edited by Gilbert Chin

PHYSICS

When Photons Bunch



Setup to probe antibunching to bunching behavior.

Being bosons, photons like to group together, with the behavior of photon bunching described as an attribute of classical light. At the other extreme, photons emitted by a single emitter are expected to antibunch, trickling out of the emitter one at a time. Although bunching and antibunching are well established behaviors of classical and nonclassical light, respectively, the transition between the two has not been observed. It is expected that as the number of emitters is increased, a smooth transition should occur. Using a high-quality cavity into which they can place a variable number of atoms, Hennrich *et al.* show that they can probe the transition systematically as the number of emitters (atoms in the cavity) is gradually increased. They observe that the antibunching behavior disappears when the average number of atoms in the cavity is one, and they are able to explain the experimental data well if the emitters are assumed to form an independent ensemble. — ISO

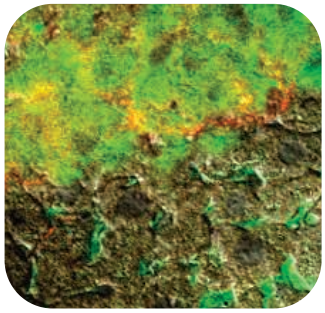
Phys. Rev. Lett. **94**, 053604 (2005).

BIOMEDICINE

Stanching the Flow

Hemophilia B is an X-linked genetic disorder caused by decreased levels of factor IX, which functions as part of the blood-clotting cascade. Deficiencies in blood clotting result in uncontrolled bleeding in response to even the slightest trauma. Another problem is bleeding into joints, where subsequent inflammation contributes to deterioration of the joint. Injecting factor IX serves as treatment to stop bleeding, and restoring a fraction of the normal amount can make a difference; however, factor IX does not survive for long in the bloodstream.

Fair *et al.* have shown in mice how embryonic stem cells can be used as a therapy for factor IX deficiency, an approach that would avoid the need for repeated injections and could supply a steady stream of factor IX. In these experiments, the mice carried a mutation in their factor IX gene, and the embryonic stem cells were derived from mice with a normal factor IX gene. In vitro culture conditions were defined to direct the embryonic stem cells to differentiate into cells with features of endodermal precursors. These putative endodermal precursors were then injected into the livers of factor IX–deficient mice. Mice treated in this way showed factor IX expression and improved long-term survival. Engraftment of the differentiated embryonic stem cells did not require injury or hepatectomy. The results provide a promising step toward a cell-based therapy for factor IX deficiency. — PJH



Liver slice showing colocalization of factor IX (red) and embryonic stem cells (green).

Proc. Natl. Acad. Sci. U.S.A. **102**, 2958 (2005).

CHEMISTRY

Beyond the Basics

The formation of carbon-carbon bonds via enolate addition to electrophiles is a cornerstone of modern organic synthesis. These types of reactions are generally started by treating a ketone with a base, which deprotonates the carbon atom adjacent to the carbonyl group, leaving a negatively charged O-C-C framework that is the target of the electrophile. However, the product is also a ketone and hence is susceptible to repeated attack by the base, leading to losses in stereoselectivity and undesirable side reactions.

Trost and Xu have found a way around this problem by eliminating the base. They stabilized precursors in the enol form by tethering the electrophile (an allyl group in this case) to the enol oxygen through a carbonate (OCO_2) linkage. Activation of the allyl group by an asymmetric palladium catalyst liberates the CO_2 spacer and allows the enol and allyl carbons to join without a deprotonation step. This reaction provides access to a broad range of tertiary and quaternary carbon centers in good yield and enantioselectivity, while minimizing side reactions. — JSY

J. Am. Chem. Soc. 10.1021/ja043472c (2005).

GEOCHEMISTRY

Elemental Traces

Microbes modify soil. What is obvious to most of us is that they process organic matter derived from higher plants and animals. What is less apparent is that they produce organic ligands and acids that bind to elements such as iron and other metals, and this affects their solubility and mobility. Thus, a soil containing microbes has a different inorganic chemistry than one lacking microbes.

The mobility of elements can be used as a measure of leaching and integrated rainfall and also as an indication of microbial activity. Neaman *et al.* have used elemental mobilities to help ascertain whether microorganisms had managed to invade Earth's land surface in the Archean. For consistency, the authors examined several ancient soils produced on one rock type—basaltic lava flows—and simulated the effects of microorganisms on such a soil in the laboratory. The presence of organic ligands greatly increased the mobility of Fe and P, changing the soil profile. These effects were evident in soils dating to 2.7 billion years ago, implying that at least some microorganisms were a significant presence on Earth's surface then, not just in the oceans. — BH

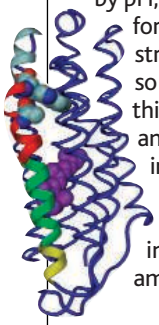
Geology **33**, 117 (2005).

CONTINUED ON PAGE 1173

BIOPHYSICS

Deconstructing Membrane Proteins

Progress in understanding how a protein finds its three-dimensional structure in seconds has been hard-won, and some of the successes have come from studying the intermediate stages (or lack thereof) of protein structures when they are stressed by pH, denaturants, or mechanical force. The historic nomenclature of structures (primary, secondary, and so forth) largely reflects the current thinking that helices form early and relatively independently, that interactions between helices help steer the folding trajectory (by clamping posts and beams) into domains, and that fitting amino acid side chains into pockets (like tenons and mortises) locks everything into place. Most unfolding studies have avoided the complications of membranes; structure determination of intact membrane proteins is not easy, and the study of pH- or denaturant-treated membrane proteins is truly daunting.



A view of helix E and the Ala-Trp interaction.

Cisneros *et al.* have applied mechanical force to extract halorhodopsin from its native membrane and compared the force-distance profiles with those of its cousin bacteriorhodopsin. The adhesive interhelical contacts are both weak and spatially diffuse, so that it is the sum total of them and not just a few residues that lend strength to the functional structure. The unfolding profiles also show, within one of the transmembrane helices in halorhodopsin, a hinge (defined by an alanine-tryptophan pairing) that demarcates two separably movable segments of the helix. — GJC

Structure **13**, 235 (2005).

BIOMEDICINE

Pockets of Resistance

Although only 5% of those exposed to mycobacteria go on to develop acute tuberculosis, many suffer latent infections that have escaped antibiotic treatment and may recrudesce with stress or aging. Ha *et al.* tested a combined vaccine-chemotherapy regime for its ability to prevent reactivation of disease in mice infected with *Mycobacterium tuberculosis*. Although protective antigens have not yet been defined precisely for tuberculosis, these authors made a DNA vaccine in a pGX10 vector containing two genes they had tested previously: Ag85A epitopes (recognized by $CD4^+$ T cells) are expressed on the surface of macrophages during early infection and PstS-3 epitopes (recognized by both $CD4^+$ and $CD8^+$ T cells) during late phase. Four weeks after infection, they administered the vaccine to mice along with the drugs isoniazid and pyrazinamide. Subsequent treatment with dexamethasone reactivated the disease in the control groups of mice but not in the vaccinated and antibiotic-dosed mice, suggesting that combining the boosting of the immune response with drugs had eliminated the mycobacteria. — CA

Gene Ther. 10.1038/sj.gt.3302465 (2005).

HIGHLIGHTED IN *SCIENCE*'S SIGNAL TRANSDUCTION KNOWLEDGE ENVIRONMENT

Science's STKE www.stke.org

PIs as Ligands?

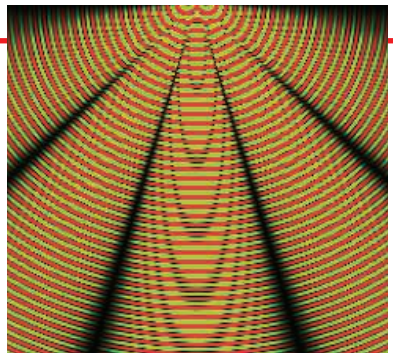
Phosphatidylinositol (PI) lipids are implicated in a broad range of processes, from the organization of signaling pathways to vesicle trafficking and control of the actin cytoskeleton. Krylova *et al.* suggest that these lipids may also serve as activating ligands for a class of orphan (so called because no regulatory ligand was known) nuclear receptors. The authors solved crystal structures of three nuclear receptor 5A family members: mouse mSF-1 and the human proteins hSF-1 and hLRH-1. Residual electron density in the ligand-binding pockets revealed that the crystallized proteins (expressed in and purified from bacteria) contained lipids. Testing with eukaryotic lipids revealed preferential binding to phosphatidylinositol 3,5-bisphosphate [PI(3,5)P2] and phosphatidylinositol 3,4,5-trisphosphate [PI(3,4,5)P3]. Although biological regulation by such lipids remains to be explored, mutant proteins designed to disrupt lipid binding showed decreased transcriptional activity. The mouse receptor appears to have lost ligand-binding activity, and phylogenetic analysis favors the scenario in which the ancestral nuclear receptor did bind lipids and this capacity was later lost in the rodent lineage. — LBR

Cell **120**, 343 (2005).

CREDITS: CISNEROS *ET AL.*, *STRUCTURE* **13**, 235 (2005)

NetWatch

edited by Mitch Leslie



EDUCATION

Physics Tutor

Desperately seeking a tutorial on fractals, a graduate-level biophysics text, or a Web demonstration of Newton's laws of motion? Visit the Physical Sciences Resource Center, a collection of mostly free educational links compiled by the American Association of Physics Teachers. The clearinghouse spans the physical science universe—from quantum mechanics to meteorology to astronomy—and includes plenty of offerings for college and graduate classes. For example, firing up the math and physics Java applets created by software developer Paul Falstad of Minneapolis, Minnesota, lets students simulate everything from the orbitals of a hydrogen atom to interference between waves passing through two slits in a barrier (above).

www.psrc-online.org

DATABASE

Genetics of Seizures

The new database CarpeDB profiles some 400 genes linked to epilepsy in humans, mice, nematodes, and other organisms. Cody Locke, an undergraduate working with neurobiologists Guy and Kim Caldwell at the University of Alabama, Tuscaloosa, compiled the collection for researchers studying the genetics of epilepsy and related conditions. Pick a gene such as *ENFL2*, which triggers nocturnal seizures, and you'll summon a virtual index card listing the gene's chromosomal location, the condition it contributes to, references, and other data. For further information, follow links to sites such as Online Mendelian Inheritance in Man and UniProt.

www.carpedb.ua.edu

TOOLS

Zooming In on SNPs

The human genome teems with SNPs, single-letter changes in DNA that can increase susceptibility to diseases such as cancer and Alzheimer's. For help viewing and tracking these variations, check out the new Genewindow from the U.S. National Cancer Institute. Genewindow is a DNA browser that lets you scroll through a particular gene nucleotide by nucleotide, identifying SNPs and other landmarks, such as protein-coding segments. You can also see how each SNP changes the sequence of the gene's protein. The tool requires the free Adobe SVG viewer and can be balky with some browsers.

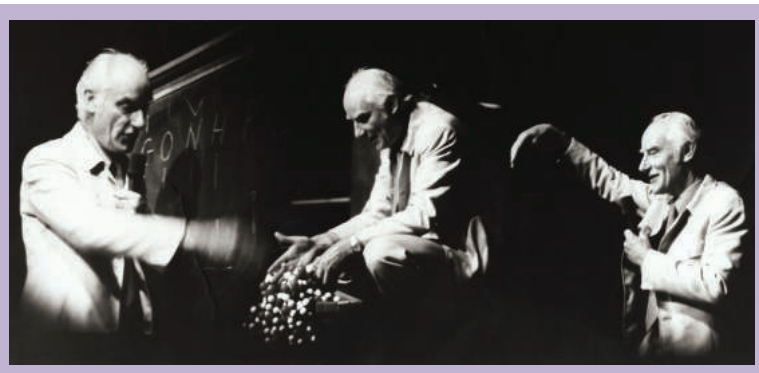
genewindow.nci.nih.gov

TOOLS

Digital Molecular Library

To uncover compounds that might jam HIV's surface molecules or block a key enzyme in cancer cells, researchers can use computer programs that predict whether particular molecules fit together. But databases of candidate compounds for virtual screening are often expensive, a limitation that inspired chemists at the University of California, San Francisco, to launch the free database ZINC. The site holds three-dimensional versions of more than 2.7 million small molecules that users can plug into common structure-matching programs. The list comes from the catalogs of 10 chemical suppliers, so you can order promising compounds. ZINC can customize sets of molecules for testing, and a virtual sketchpad lets you specify substances that carry a particular chemical group.

blaster.docking.org/zinc



EXHIBIT

After the Double Helix

"I have never seen Francis Crick in a modest mood," James Watson declared in his controversial book *The Double Helix*. Regardless of whether the characterization was accurate, Crick (1916–2004) had plenty to be immodest about. As you can see at this new exhibit on his life from the U.S. National Library of Medicine, Crick's contributions went far beyond co-discovering the structure of DNA.

After helping set the research agenda for molecular biology's early years, Crick at age 60 launched a new career as a neuroscientist, theorizing about questions such as the origin of consciousness and the function of rapid eye movement sleep. Along with a biography that follows his professional zigzags, the site holds letters, papers, photos, and other memorabilia from a collection cached at the U.K.'s Wellcome Library. You can peruse an early sketch of the double helix, for example, or read a letter from chemist Linus Pauling chastising Crick for including too few hydrogen bonds in a paper on DNA. This 1979 composite photo (above) shows Crick's animated lecture style.

profiles.nlm.nih.gov/SC

CREDITS (TOP TO BOTTOM): PAUL FALSTAD; COURTESY OF THE WELLCOME LIBRARY

Send site suggestions to netwatch@aaas.org. Archive: www.sciencemag.org/netwatch

Published by AAAS

This Week

PAGE 1182
Global view of global observations

1185
Dispute over HIV "superstrain"

ASTROPHYSICS

Giant Neutron-Star Flare Blitzes The Galaxy With Gamma Rays

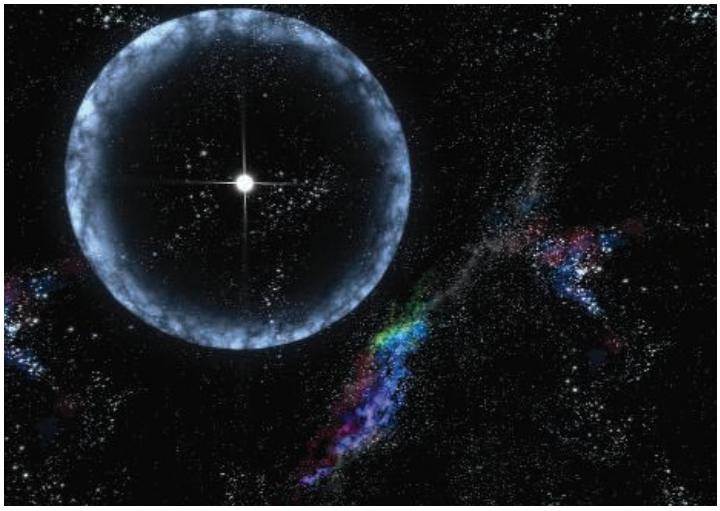
WASHINGTON, D.C.—Starstruck astrophysicists are agog over an explosion toward the center of our galaxy that irradiated Earth with gamma rays and x-rays shortly after Christmas. The outburst was brighter than any solar eruption ever measured, even from its estimated distance of tens of thousands of light-years. The source, scientists believe, was the most exotic kind of neutron star: a "magnetar" shot through with twisted magnetic fields, powerful enough to shred the ultradense surface of the tiny object.

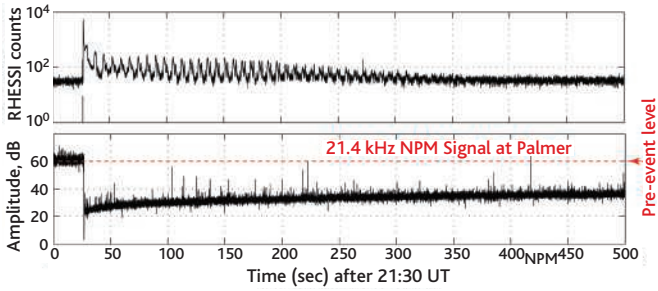
Lasting just one-fifth of a second, the flare released as much energy as our sun produces in 250,000 years. It spawned a fireball that radio telescopes see blasting into space at 30% of the speed of light. That pattern—an intense spike of energy and a fading afterglow—leads researchers to suspect that they caught a miniature gamma ray burst (GRB), startlingly close to home.

Although the biggest GRBs probably arise from giant stars that collapse into black holes in remote galaxies, magnetar flares in nearby galaxies might produce some of the short bursts that no one could explain until now. "For all practical purposes, this could be a short GRB in our own galaxy," says astrophysicist Chryssa Kouveliotou of NASA's Marshall Space Flight Center in Huntsville, Alabama. "It's an amazing event, something I never expected to see in my lifetime."

Other scientists used similar superlatives at a NASA press conference here on 18 February, convened far in advance of several reports about the flare to appear in *Nature*. NASA officials fretted that the story would elude their control, thanks to other papers already posted online. The hasty publicity alienated members of some teams, who felt excluded until NASA agreed to cite their work during the briefing.

But the flare's glow should erase those hard feelings as observers and theorists race to understand its origins. The fierce shell of radiation tripped detectors on about 15 satellites and solar-system probes on 27 December, says astronomer Kevin Hurley of the University of California, Berkeley. Scientists quickly traced the flare to a magnetar called SGR 1806-20, in a crowded and dusty part of the sky toward the Milky Way's center.





Sizzling shell. Radiation floods the galaxy from an ultramagnetic neutron star (*above*). The intense pulse (*right*, top graph) ionized Earth's daytime atmosphere to startlingly low levels (bottom graph).

Of the dozen or so suspected magnetars astronomers have identified so far, SGR 1806-20 is the third to spew a giant flare seen by modern telescopes. The two previous outbursts, in 1979 and 1998, helped researchers devise their theory that SGR 1806-20 and its kin are slowly rotating neutron stars powered by the most intense magnetic fields known (*Science*, 23 April 2004, p. 534). The latest flare—about 100 times stronger than the others—probably arose from a sudden transformation of the entire neutron star, theorists believe.

The flare's impact on Earth was stunning. It struck during the daytime over the Pacific Ocean, from a spot on the sky just 5 angular degrees from the sun. Very-low-frequency radio transmissions between Hawaii and Antarctica showed that Earth's ionosphere compressed tens of kilometers inward from its usual daytime altitude of about 70 kilometers, says atmospheric physicist Umran Inan of Stanford University in California. Moreover, the disturbance lingered for an hour. "That's totally unheard of," says Inan, who thinks the radiation altered chemical reactions that typically restore the ionosphere after solar flares.

"I'm awestruck by that," says Dale Frail of the National Radio Astronomy Observatory in Socorro, New Mexico. "This little thing near the center of our galaxy reached out and physically invaded us, even though it's just 20 kilometers across."

The flare's position near the sun prevented most telescopes and satellites from taking a close look. But radio dishes across the globe—immune to sunlight—took aim in the days after. Most notably, two teams used the Very Large Array of 27 radio telescopes in Socorro to look for a glowing blast wave. "All hell broke loose seconds after the data came in," says Bryan Gaensler of the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts. "[The afterglow] was 500 times brighter than we expected. I thought we had pointed the telescope at the wrong place."

Meanwhile, astrophysicists worked furiously to figure out the flare's actual energy, a tough task because most radiation detectors were so swamped. For example, NASA's Swift satellite—launched in November to study GRBs—was overwhelmed by nearly a trillion photons that pierced the body of the spacecraft, says astrophysicist Neil Gehrels of NASA's Goddard Space Flight Center in Greenbelt, Maryland. But fingernail-sized particle detectors on a few satellites kept up with the onslaught, enabling researchers to calibrate the flare's 0.2-second fury.

Two independent teams, led by Hurley ▶

CREDITS (TOP TO BOTTOM): NASA; UMRAN INAN/STANFORD UNIVERSITY

Published by AAAS

and by astrophysicist David Palmer of Los Alamos National Laboratory in New Mexico, conclude in their upcoming *Nature* papers that the energy flux resembles the spikes of radiation seen from short GRBs in other galaxies. But opinions differ over just how many magnetar flares pop off within about 300 million light-years from the Milky Way, the expected viewing range of satellites. "This type of event conceivably could explain all of [the short GRBs]," says Palmer. Others argue that most short GRBs still must arise from different sources—colliding neutron stars, for instance.

Debate also rages about the physics that powered the flare. Four years ago, David Eichler of Ben-Gurion University in Beer-Sheva, Israel, forecast that a wholesale decay of magnetic fields outside a magnetar could spark a supergiant flare. Magnetar theorists including Robert Duncan of the University of Texas, Austin, now propose a variation: The flare arose from wound-up magnetic fields deep inside the neutron star that catastrophically "untwisted" near and above the surface. As the fields snapped into new configurations, the theorists believe, they sheared the entire surface of the star and unleashed a 2-billion-degree fireball of electrons and positrons.

As SGR 1806-20 moves safely away from the sun, more observatories will examine the fading outburst—including infrared telescopes on the ground and x-ray telescopes in space. Astrophysicists eagerly await each result, creating a tumult that the field has not seen since Supernova 1987A flared into view 18 years ago this month. "I heard stories that the supernova was a life-changing experience for many astronomers," says Gaensler. "I now understand why."

–ROBERT IRION

HUMAN ORIGINS

Battle Erupts Over the 'Hobbit' Bones

Research on human fossils generally proceeds at a leisurely pace. Those who discover new bones sometimes take years to analyze them, while their colleagues and rivals wait impatiently to get a good look. But that's not the case with the 18,000-year-old "hobbit" skeleton of Indonesia. Ever since the Australian-Indonesian team that discovered the bones made the startling claim that they are the remains of a species of small, archaic human, *Homo floresiensis* (*Nature*, 28 October, p. 1055), the bones have been analyzed and reanalyzed at a breathtaking pace. For the past 3 months, however, the studies have been directed not by the discoverers but by a rival who has taken possession of the skeleton.

The bones were discovered in 2003 in the Liang Bua cave on the Indonesian island of Flores by a team led by archaeologist Mike Morwood of the University of New England, Armidale. But in November last year, the Center for Archaeology in Jakarta agreed to let Indonesian paleoanthropologist Teuku Jacob study the skeleton in his laboratory at Gadjah Mada University in Yogyakarta (*Science*, 12 November 2004, p. 1116). Jacob has since invited other researchers to inspect it and, in a move that Morwood calls "unethical" and "illegal," asked a team from the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, to conduct DNA and other analyses on a 1-gram sliver of rib. Jean-Jacques Hublin, director of the department of human evolution at the Max Planck, who carried the sample to Leipzig, counters that he has "formal authorization" from Tony Djubiantono, head of the Jakarta archaeology center, to analyze the sample.

Jacob has announced that, in his view, the skeleton is that of a modern human pygmy with microcephaly. "There are [living] pygmies near there—near Liang Bua," he notes. Last week, three paleoanthropologists, two of whom had publicly challenged Morwood and his colleagues' analysis—Maciej Henneberg of the University of Adelaide, Alan Thorne of Australian National University in Canberra, and Robert Eckhardt of Pennsylvania State University—announced, after examining the bones with an Australian camera crew looking on, that they agree with Jacob. Morwood calls this interpretation "mind-boggling."

DNA analysis could settle the case. If "the DNA sequence falls outside the variability seen in modern human sequences, then we could be confident that it's not a modern human," says Svante Pääbo of the Max Planck. But he estimates that the sample has a "less than 50%" chance of yielding ancient DNA. And "if it carries a sequence similar to a modern human, we will not be able to exclude contamination [with DNA from those who have handled the sample] completely," says Pääbo.

Max Planck scientists also may do stable isotope analysis, which offers clues to diet, and radiocarbon dating. "Professor Jacob has the fullest authority to ask us to perform this analysis," says Hublin, who has worked with Jacob for years.

Bone of contention. Teuku Jacob has commissioned DNA analysis of the skeleton.

Behind the infighting lies the question of who should have the right to study important fossils. Jacob argues that for decades, archaeologists have brought bones excavated from Liang Bua to his laboratory for anatomical analyses. But Morwood points out that he has a signed agreement with the archaeological center, the official repository for the bones, allowing his team to work with the center's scientists to study the specimen. "We haven't been able to do the analysis we'd want to do because we haven't seen the specimen since [Jacob] took it in November," says geologist and Morwood colleague Bert Roberts of the University of Wollongong. The deadline for returning the bones was 1 January, but Djubiantono has extended it twice.

Last week Jacob told *Science* that he had "almost finished" analyzing the specimen. But Morwood, speaking from the field in East Java, says he is not optimistic about the bones' return to Jakarta. "We're moving on with our research," he says, planning publications based on previous measurements and also mounting new field expeditions to gather more material. Jacob says he, too, plans "a series of articles on different aspects, with colleagues." The conflict continues—and for now, at least, so does the research.

–ELIZABETH CULOTTA

CREDIT: AGUS SUPARTO/AFP/GETTY IMAGES

SOUTH KOREA

Radical Reforms Would Shake Up Leading Science Institute

SEOUL—The Korean government was looking for fresh ideas when it hired physics Nobelist Robert Laughlin last summer as president of the Korea Advanced Institute of Science and Technology (KAIST), one of the country's top science and engineering universities. It may have gotten more than it bargained for.

Laughlin has floated a plan to cure the prestigious institution of its "addiction" to government subsidies. His prescription—more undergraduates, higher tuition, and courses that appeal to nonscience majors—has, however, been loudly denounced by some faculty members as a danger to the patient.

CREDIT: COURTESY OF KAIST

Strong medicine. Robert Laughlin says KAIST needs to end its "addiction" to government funding.

"KAIST is too old and too important to be the experiment of one person's ridiculous ideas," says Park O-ok, who resigned as dean of office planning after Laughlin circulated a draft of his plan in December. Others say the proposal is simply unworkable. "I'm open-minded about it," said one member of the faculty committee that reviewed the plan, which is expected to be formally unveiled on 1 March. "But he has no chance of succeeding."

Established in 1971, KAIST was intended to provide essential talent for Korea's high-tech sector. But the number of students interested in scientific careers has declined steadily since the mid-1990s, and government officials thought that reforming KAIST might help reverse the trend. Last summer they took such a step by hiring Laughlin, a physics professor at Stanford University whose explanation of how electrons acting together in strong magnetic fields can form new types of "particles" with fractional charges earned him a Nobel Prize in 1998 (*Science*, 4 June 2004, p. 1427). He was the first foreigner to lead a university or major government research institution in Korea.

Laughlin wasted no time. Within 6 months he had decided that KAIST was being "squeezed [financially], with no exit." The solution, he declared, was to increase income from sources other than the government. In December, he sent a draft plan to the Ministry of Science and Technology and a faculty committee that would triple KAIST's current enrollment of 7300 and shift the balance toward undergraduates, quadruple the current $850-a-semester tuition, revamp the undergraduate curriculum by appealing to premed and prelaw students, and tweak its graduate research programs to turn a profit.

"I do not think that the problem of lack of interest in science should be solved," he said provocatively at his inauguration. "The right course of action is to change science and technology so it becomes economically competitive." In other words, reduce the number of poorly enrolled basic research classes in favor of subjects that will attract students.

Laughlin holds up the Massachusetts Institute of Technology as the model of an institution attuned to finding outside sources of revenue. "KAIST was made to subsidize industry," he told *The Korea Times* in January. "MIT was set up to make money."

Critics say that Laughlin has ignored an existing 10-year plan aimed at achieving financial independence. "The idea of getting an endowment and finding industry financing was discussed a long time ago," says physicist Shin Jung-hoon, a member of the 21-person faculty review committee. His blunt, forceful style, they add, has bruised egos in the face-conscious country. Although Laughlin says he clearly labeled his December draft as "the starting point for discussion … with strong language to sharpen the issues," Park and others say that caveat wasn't on the original documents.

Despite the controversy, Laughlin remains optimistic that his vision will eventually prevail. "They're being terrific," he said of the KAIST faculty and the science ministry, which oversees KAIST. "This is an unusual and historic cultural experiment. Of course it isn't easy."

—MARK RUSSELL

Mark Russell is a freelance writer in Seoul.

ScienceScope

A Budget Bouillabaisse

Several U.S. science agencies will have new congressional budget bosses this year after the House of Representatives last week rearranged its spending panels in hopes of streamlining operations. If the Senate doesn't follow suit, however, the misalignment could result in a more chaotic 2006 budget cycle.

Compressing the current 13 House spending panels into 10 throws the National Science Foundation (NSF) and NASA into a polyglot subcommittee, chaired by Representative Frank Wolf (R–VA), that already oversees two Commerce Department science agencies: the National Oceanic and Atmospheric Administration and the National Institute of Standards and Technology. The Environmental Protection Agency joins a panel that includes the Interior Department's U.S. Geological Survey. The National Institutes of Health is unaffected by the reshuffling.

Senator Christopher Bond (R–MO) could lose his gavel if the Senate follows suit. And last week, at a hearing on NSF's budget, he warned scientists that "basic research will suffer" under a catch-all spending bill, the likely product of a Congress without parallel spending panels.

—JEFFREY MERVIS

Fiscal Woes Dog Gamma Ray Satellite

Budgets are more dangerous than gamma ray bursts are, at least for the Gamma Ray Large Area Space Telescope (GLAST) project.

GLAST, a $685 million satellite sponsored by NASA, the Department of Energy, and a number of foreign partners, is intended to pick up traces of gamma ray bursts and other violent astrophysical phenomena. But it's running perhaps $25 million over budget because of problems in building and developing 16 detectors at the heart of the telescope—devices that measure the direction and energy of incoming gamma rays.

A NASA review will decide whether to proceed as planned or eliminate some of the 16 detectors. The latter choice would save only a small percentage of the satellite's bill and have a serious scientific cost, notes Naval Research Laboratory astrophysicist Charles Dermer. "You may lose new classes of [gamma ray] sources as a consequence," he says. "It's not a good payoff."

—CHARLES SEIFE

Published by AAAS

NATIONAL INSTITUTES OF HEALTH

NCI Gears Up for Cancer Genome Project

The National Cancer Institute (NCI) is hoping to launch a $1.5 billion effort to identify all major mutations in the most common human cancers. The 10-year project would gather tumor samples from thousands of patients and scrutinize them for genetic glitches. "If we can sequence the cancer genome, … I think we must," says Anna Barker, NCI deputy director for strategic scientific initiatives.

The idea for a Human Cancer Genome Project comes from a working group of the National Cancer Advisory Board (NCAB) commissioned by NCI Director Andrew von Eschenbach to find new opportunities to use technologies in cancer. The panel, led by Eric Lander of the Massachusetts Institute of Technology Broad Institute in Cambridge and Lee Hartwell of the Fred Hutchinson Cancer Research Center in Seattle, Washington, spent 18 months hammering out a plan, which includes expanding NCI's biomarker efforts (*Science*, 12 November 2004, p. 1119). The cancer genome is the one "specific project" they came up with, Lander told NCAB last week.

Insights into the genetic basis of cancer have led to targeted drugs like Gleevec, Herceptin, and Iressa, noted Lander. "But we know a minority of the story" about genetic changes in tumors. A "comprehensive" effort to identify these mutations, he added, "would propel the work of thousands of investigators."

As outlined in a 21-page white paper, the project would collect tumor samples from patient volunteers and use technologies such as gene chips to find mutated regions. Those sections would then be resequenced to identify specific mutations. The goal—to identify all mutations occurring at 5% frequency in the 50 most common types of cancer—would require 250 samples per type, or 15,000 samples. The work would be carried out by "sample acquisition centers" and genome analysis centers, and it would include studies of ethical and intellectual-property issues. It would begin with roughly $50 million a year in pilot projects for a few years before ramping up to $200 million a year.

The project would be jointly managed by NCI and the National Human Genome Research Institute (NHGRI), which is already on board. "I am extremely enthusiastic about this proposal," NHGRI Director Francis Collins told NCAB. "The potential value here is massive," said cancer board member Franklyn Prendergast of the Mayo Clinic in Rochester, Minnesota; another member called it "mind-boggling."

Advocates hope that the price tag is not. Although Lander predicts that the project would "require new funds" from Congress and perhaps industry, his working group wants NCI to launch the pilot projects next year. To do that, von Eschenbach has asked the board to begin prioritizing other programs that can be cut back or eliminated. Although von Eschenbach didn't offer specific suggestions, he noted that "we are applying a lot of fiscal discipline within our own house." **–JOCELYN KAISER**

ENVIRONMENTAL SCIENCE

Forging a Global Network to Watch the Planet

CAMBRIDGE, U.K.—The dream of creating a global earth-monitoring network came a step closer to reality last week. Proponents met in Brussels to launch a 10-year program to turn gauges, sensors, buoys, weather stations, and satellites that monitor Earth's surface, atmosphere, and oceans into a unified whole. The Global Earth Observation System of Systems (GEOSS), as it's called, is expected to evolve slowly from national systems into a comprehensive, coordinated, and sustained set of observations for the benefit of everyone, including developing countries. By adding links and standards, "earth science will step up to the next level: a total earth-observing system," says Conrad Lautenbacher Jr., head of the U.S. National Oceanic and Atmospheric Administration.

There is little coordination today among the roughly 50 satellites observing Earth—or the more numerous sensors in ground- and ocean-based networks. As a result, there are gaps in coverage as well as a massive duplication of effort. The drive to add coherence began in July 2003 when government ministers from some 30 nations along with heads of various agencies—collectively dubbed the Group on Earth Observations (GEO)—met for an Earth observation summit in Washington, D.C. At a second summit in Tokyo in April 2004, GEO came up with the idea of having a 10-year transition from the current hodgepodge of observations to a global coordinated system. And on 16 February in Brussels, the third summit signed off on GEOSS's 10-year implementation plan. The agreement puts GEO itself—with 60 countries and 33 organizations now on board—on concrete footing, with a permanent secretariat hosted by the U.N.'s World Meteorological Organization in Geneva.

Center of attention. Brussels, shown in this 2003 image from a European environmental satellite, hosted a summit to launch a decadal plan for GEOSS.

CREDIT: 2004 ESA

The goals of GEOSS are lofty: Its proponents say it will improve weather forecasts, reduce the devastation of natural disasters, monitor climate change, support sustainable agriculture, help understand the effect of environment on human health, and protect and manage water and energy resources.

The first job, according to Errol Levy, scientific officer at the European Commission in Brussels, "is to look at what is measured now, what is needed," and find the gaps. GEOSS will initially build on existing satellites and sensors, such as NASA's Earth Observing System satellites and the European Space Agency's Envisat. The most difficult part, Lautenbacher says, will be achieving an agreement on data sharing. There will have to be some horse trading on who observes what and who launches which satellite. "Some of this is difficult. It will require serious negotiation," Lautenbacher says.

One of GEOSS's key aims is to involve developing countries. Lautenbacher says they "have the most to gain," in terms of helping them tackle problems such as desertification or the spread of malaria. But they can also contribute by launching weather balloons into the upper atmosphere or installing tide gauges to measure sea-level rise. **–DANIEL CLERY**

Published by AAAS

DRUG SAFETY

FDA Panel Urges Caution on Many Anti-Inflammatory Drugs

GAITHERSBURG, MARYLAND—At a 3-day meeting last week to hash out how a class of painkillers might trigger heart attacks and strokes, new puzzles emerged, even as two U.S. Food and Drug Administration (FDA) advisory committees agreed that use of the drugs, known as COX-2 inhibitors, be allowed but sharply curtailed. Wrestling with questions that affect millions of patients and billions of dollars in drug company revenue, the committees—on arthritis drugs and drug safety—voted in favor of keeping Celebrex, Bextra, and even Vioxx on the market but adding stringent warnings to them and possibly other nonsteroidal anti-inflammatory drugs (NSAIDs). Steven Galson, acting head of FDA's Center for Drug Evaluation and Research, said in advance that FDA would act rapidly on the panel's recommendations, likely within a few weeks.

Tough choices. Naproxen, marketed as Aleve, got more backing for its safety profile than either Celebrex or diclofenac, marketed as Voltaren.

But the recommendations belie a confusion that, if anything, has intensified in the 5 months since Merck withdrew its COX-2 inhibitor Vioxx from the market. At issue are two questions, neither of which could be definitively answered last week, despite exhaustive parsing of clinical trials data involving tens of thousands of volunteers. How do drugs that inhibit COX-2, which mediates inflammation, disrupt the cardiovascular system? And which drugs present a significant risk?

"We need to worry about the data we don't have," said James Witter, an FDA rheumatologist who reviewed Pfizer's Celebrex and Bextra, the COX-2 inhibitors that remain on the U.S. market.

COX-2 drugs are no more effective at controlling pain than traditional NSAIDs like naproxen, but they quickly gained favor because they don't cause NSAID-associated stomach problems. The reason is that COX-2 inhibitors can blunt COX-2 while steering mostly clear of COX-1, a related enzyme. Dampening COX-1 can cause stomach ulcers. Traditional NSAIDs as well as COX-2 drugs may also protect against and help treat cancer.

Merck voluntarily withdrew Vioxx after a colon cancer prevention trial suggested that it nearly doubled the rate of strokes and heart attacks compared with a placebo. At first, many scientists believed that Vioxx's unusually selective targeting of COX-2 was the culprit. That, they thought, might be upsetting a fine balance between two fatty acids, prostacyclin and thromboxane, that control blood clotting. By that reasoning, related drugs like Celebrex, a slightly less targeted COX-2 inhibitor with a shorter half-life, as well as traditional NSAIDs, might be safe, or at least safer. Even when a 2000-person Celebrex trial was halted in December after participants taking the drug suffered roughly two to three times more cardiovascular problems, the targeting theory remained plausible.

New data, however, hint that something else might be at work, said committee member Steven Abramson, a rheumatologist at New York University and the Hospital for Joint Diseases in New York City. He and others were struck by a Pfizer study of cardiac surgery patients posted online last week in the *New England Journal of Medicine*. That study, which Kenneth Verburg, head of the company's arthritis drugs division, described at the meeting, found that participants taking Bextra reported nearly three times the rate of cardiovascular events compared with those on placebo.

The puzzle for Abramson was that all the study volunteers were also taking aspirin—commonly used by cardiac patients and by those with arthritis. Because aspirin inhibits COX-1 as well as COX-2, noted Abramson, it might be expected to counteract the cardiac risks caused by inhibiting COX-2 alone. That didn't appear to be the case in the Bextra study. "We're left with a lot of unknowns" about why Bextra caused problems despite aspirin, said Verburg.

If simply blocking COX-2, regardless of what else is also blocked, is at least partly to blame, then a much broader class of drugs—the traditional NSAIDs—could be implicated. "I think you have to tell people there may be a problem here," said panel member and cardiologist Steven Nissen of the Cleveland Clinic. The group voted unanimously to ▶

CREDITS (CLOCKWISE FROM TOP LEFT): STEVEN SENNE/AP PHOTO; MARY ALTAFFER/AP PHOTO; NOVARTIS PHARMACEUTICALS

ScienceScope

Huge HIV Vaccine Gift From Gates

The Bill and Melinda Gates Foundation once again has put its money where its mouth is, with a $360 million investment in AIDS vaccine research.

The money, so far the largest single contribution to the field, supports the goals of the HIV/AIDS Global Vaccine Enterprise, an international alliance that the Gates Foundation helped organize. The new funds will go to research centers or consortia that are trying to improve tests to assess the immune responses triggered by vaccines or design vaccines that stimulate either antibodies or cell-mediated immunity. Letters of inquiry are due by 1 April. The size and number of awards will depend upon the quality and scope of the proposals. —JON COHEN

Will Stem Cell Research Restrictions Be Lifted?

Emboldened by what they say is overwhelming public support, federal legislators from both parties have introduced a bill in the House to allow federal funding for research on stem cell lines derived after August 2001 from leftover zygotes in fertility clinics. In an apparent bid to placate opponents, the legislation would not sanction government funding to derive the lines.

Although a similar push stalled last year, Representative Diana DeGette (D–CO) says the House bill (H.R. 810) already has 190 cosponsors, and she promises "to use every legislative [means] to bring it up." An identical Senate version is expected shortly.

In other news, on 18 February, the United Nations' legal committee approved a nonbinding resolution urging states to ban all forms of human cloning.

—CONSTANCE HOLDEN

Pasteur Move Not Needed, Mediator Says

PARIS—A British mediator has sided with French scientists fighting a relocation plan by the Pasteur Institute in Paris. The institute sought to move part of its lab to the suburbs while Pasteur's aging downtown campus is being renovated (*Science*, 21 January, p. 333).

John Skehel, director of the Medical Research Council's National Institute for Medical Research in London, told department directors last week that a phased renovation would make the relocation unnecessary.

The mediator's report is a setback for embattled Pasteur director Philippe Kourilsky, who was unavailable for comment. "There's a feeling on campus he might step down," says Pasteur virologist Simon Wain-Hobson. —MARTIN ENSERINK

Published by AAAS

add new warning labels to NSAIDs, although it cautioned that some clearly pose a greater risk than others.

That was underscored by a presentation made by FDA drug safety officer David Graham. He described his preliminary results from an epidemiologic study of Medicaid patients; those taking Indocin (generically, endomethacin) had nearly double the number of heart attacks, while those on Mobic (generically, meloxicam) had a 37% increased risk. Many patients switched to these drugs, which are thought to block COX-2 more selectively than naproxen but less selectively than Vioxx, when concerns arose about COX-2 inhibitors.

Ernest Hawk, a chemoprevention expert at the National Cancer Institute in Bethesda, Maryland, believes that NSAIDs, including COX-2 inhibitors, still show great promise as anticancer agents. "COX-2 remains a relevant oncology target," he says, backed by dozens of animal and some human studies.

A critical question, Hawk and others agreed, is whether genetic differences among patients might offer clues to their response to these drugs. Garret FitzGerald, a pharmacologist and cardiologist at the University of Pennsylvania in Philadelphia, for one, emphasized how badly such information is needed. "We've talked about personalized medicine for a long time," he said. "Here is a situation where we actually have to care" about it.

—JENNIFER COUZIN

HIV/AIDS

Experts Question Danger of 'AIDS Superbug'

"NEW AIDS SUPER BUG: Nightmare strain shows up in city," blared the 12 February headline on the front page of the *New York Post*, a tabloid not known for understatement. Even the sober *New York Times* played the story on the front page, albeit with characteristic reserve. "The story *is* the story," says AIDS researcher Steven Wolinsky, who heads the infectious disease department at Northwestern University Medical School in Chicago, Illinois.

The hubbub, which shows no sign of flagging, began on 11 February, when the New York City Department of Health and Mental Hygiene issued an alert to physicians and held a press conference, warning that a highly promiscuous New York City man was infected with a strain of HIV that was both very virulent and "difficult or impossible to treat," as it was resistant to three of four classes of anti-HIV drugs. The message was clear: Because of one man's crystal methamphetamine use and unprotected anal sex with many partners, a new killer bug might be on the loose.

But immediately AIDS researchers around the world began to question just how new or dangerous this supposed superbug really is. One of them, Julio Montaner, acting director of the B.C. Centre for Excellence in HIV/AIDS in Vancouver, co-authored a similar report in the May 2003 issue of the journal *AIDS* about two patients who became infected by drug-resistant viruses that rapidly destroyed their immune systems. Susan Little, Douglas Richman, and co-workers at the University of California, San Diego (UCSD), reported 3 years ago in the *New England Journal of Medicine* that 10% of recently infected people have mutations in the virus that are associated with resistance to at least two classes of drugs. Several researchers have also questioned just how drug resistant the man's virus is; viruses deemed "resistant" in lab tests sometimes still respond to drugs.

What's more, many researchers doubt that a multidrug-resistant strain of HIV can transmit efficiently from one person to another. "There's always a fitness cost for mutations, either functionally or structurally," explains Northwestern's Wolinsky. A study of 220 recently infected people bears this out: As Bernard Hirschel and Luc Perrin of the University of Geneva and colleagues reported in the June 2004 issue of *Antiviral Therapy*, multidrug-resistant strains were indeed less likely to transmit than "wild-type" strains.

Overblown? Critics say the New York Health Department overreacted to a single case.

But David Ho, director of the Aaron Diamond AIDS Research Center in New York City, whose lab first detected this patient's multidrug-resistant strain, stresses that this case *is* unique. Ho alerted the New York City health department after his clinic saw the patient in January. The man had last tested negative for HIV 20 months earlier and may have become infected as late as October. By the time he came to Aaron Diamond, his CD4 white blood cells, which HIV targets and destroys, had plummeted to below 30. (AIDS is defined as 200 or fewer CD4s.) Typically, it takes 10 years to progress from infection to AIDS. Ho then scoured a national database of recently infected people and looked at their CD4 loss. "There's no case like this," says Ho.

One of a kind. David Ho insists the case is unique.

UCSD's Richman downplays notions of a superbug and instead suggests that the patient may be genetically susceptible to infection or rapid development of AIDS. But Ho says a preliminary analysis of the man's genetics hasn't "found anything unusual yet."

Ho also says this strain replicates in test tube experiments slightly better than the standard reference strain used by his lab. "It's not a wimpy virus," says Ho. In fact, Aaron Diamond's Viviana Simon, Martin Markowitz, and co-workers in a July 2003 *Journal of Virology* paper showed that some drug-resistant isolates from recently infected people remain highly infectious and replicate well, suggesting that they had continued to evolve and compensated for the fitness cost the original mutations extracted.

Several researchers agree that it is premature to dub this man's HIV strain a "supervirus," as studies have yet to demonstrate that it's highly pathogenic. And many experts question whether the case deserves widespread attention in the absence of data linking the man's infection to others. "Until there's evidence that this virus is highly transmissible or even normally transmissible, it remains an anecdote," says AIDS researcher John Moore of Weill Medical College of Cornell University in New York City.

Ho says he and his co-workers hope to publish their findings rapidly online, and they will present data from this case at a large U.S. AIDS meeting in Boston this week, where those in the field can discuss it in context.

—JON COHEN

CREDITS (TOP TO BOTTOM): *NEW YORK POST*; AARON DIAMOND AIDS RESEARCH CENTER

Published by AAAS

The fight is on to save Mesopotamia's drained marshes. But it's not easy finding a realistic and salable plan—or gathering data in a dangerous environment

Reviving Iraq's Wetlands

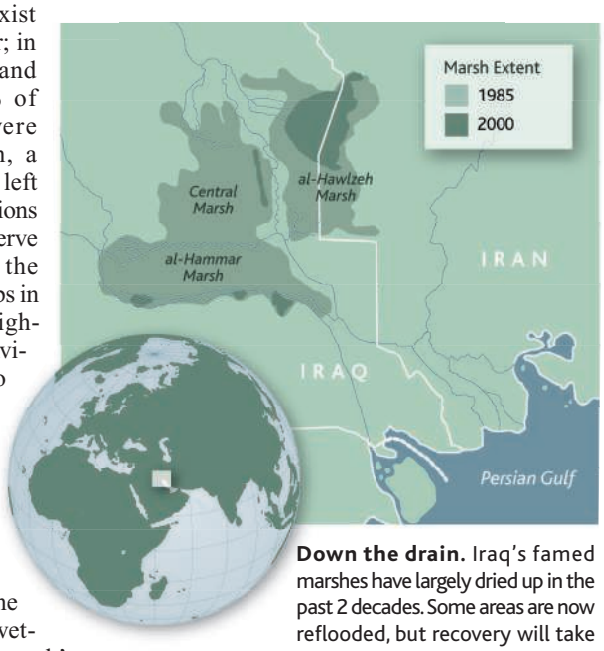
"Redeeming a swamp … comes pretty near to making a world."
—Henry David Thoreau

Azzam Alwash enjoys kayaking with his wife in southern California. But his real dream is to paddle among the high reeds of Mesopotamia's ancient marshes near where he was born. Those marshes exist mostly in his memory, however; in an unprecedented ecological and human disaster, some 90% of the famed Iraqi wetlands were destroyed by 2000. Alwash, a 49-year-old civil engineer who left Iraq a quarter-century ago, envisions a full restoration that would preserve both the vibrant wildlife and the unique culture of the Marsh Arabs in the region. He even quit his high-paying job as a partner in an environmental consulting firm to drum up international support for his effort, which he grandly dubbed Eden Again.

Alwash has helped energize a coterie of donors, scientists, local leaders, and politicians who are hotly debating the future of the marshes. The first scientific studies of the wetlands in decades appear in this week's issue of *Science* (see p. 1307), and foreign nations have pledged a total of $30 million. The Iraqi government recently set up an interagency center to draw a blueprint for revitalizing this desperately poor and ecologically battered area. But coming up with a common vision—and financing—in an unstable nation may prove even harder than collecting data. "This is a scientifically difficult and tremendously complex effort," says Edwin Theriot, a U.S. Army Corps of Engineers official who has advised the Iraqi government. "We're having difficulties with the Everglades and in Louisiana—and we're supposed to have all the resources we need."

The sheer scale of the destruction is of biblical proportions. In one generation, some 20,000 square kilometers of marsh shrank to a tenth of that size, as did a population that once numbered a half-million. Three wars and one insurrection played a big role, as did a concerted effort in the 1990s by Saddam Hussein to drain the marshes.

As the marshes turned to desert, local peoples fled or were forced from their homes. Left behind were vast salt flats laced with insecticides and landmines. The fisheries—which provided a large share of Iraq's overall catch—crashed, while animals from the Goliath heron to the pygmy cormorant face extinction.



Down the drain. Iraq's famed marshes have largely dried up in the past 2 decades. Some areas are now reflooded, but recovery will take time—and require lots of water.

The effects of the destruction radiate far beyond southern Iraq. No longer cleansed by the marshes, the salty and polluted waters flowing into the Persian Gulf from the Tigris and Euphrates rivers are playing havoc with marine life there, including the lucrative shrimp business. And Asian migratory birds have lost a major staging and wintering area on the western Siberian-Caspian-Nile flyway. "The impact on biodiversity has also been catastrophic," states a 2004 United Nations study on the marshes.

Out of Eden

Few places on Earth have a stronger hold on the imagination than do the Iraq marshes. They are the legendary site of the Garden of Eden and incubator for the first great urban centers, home of the world's first writing system. Its trackless stretches have long hidden both wildlife and rebels. Sumerian princes hunted game there, and Assyrian King Sennacherib led a force into the region in the 7th century B.C.E. to flush out pesky Chaldean rebels.

Isolated, yes, but far from pristine. "This is the oldest and most tinkered-with landscape on Earth," says Iraq's new water resources minister Abdul Latif. For at least 5000 years, humans have widened and dredged channels, dried and flooded fields, and built reed houses atop artificial islands of reed bundles. Its lifeblood was the spring floods. "This pulse of sweet fresh water, laden with sediments, flushes the salt, provides nutrients to revitalize the reed beds, and is key to bird migration," says Alwash.

Most of that water comes from outside Iraq. The Euphrates and Tigris originate in the eastern mountains of Turkey. More than 90% of the water from the Euphrates comes from Turkey, Syria, and Saudi Arabia; the Tigris's basin covers large parts of Turkey, Iran, and a slice of Syria.

Beginning in the 1950s, governments began diverting that flow, first by creating natural lakes within Iraq and later by building large dams on both rivers. There are now nearly three dozen major dams, with eight more under construction and a dozen in the planning stages. Turkey alone can store up to 91 billion cubic meters of water and will need more to irrigate its dry eastern provinces. Iraq and Syria can store as much as 23 billion cubic meters. The 2003 war and its aftermath halted plans to build additional dams in Iraq—there are currently a dozen large ones—but Iran recently embarked on a major dam-building effort on tributaries of the Tigris.

The result of this half-century of water management has been dramatic. The spring

CREDITS (TOP TO BOTTOM): RICHARDSON *ET AL.*; SOURCE: AMAR

Published by AAAS

flood is barely noticeable. The maximum flow of the Euphrates during May has dropped by two-thirds since 1974, when dam building began in earnest. Even before the 1991 Gulf War, many experts feared the result would irreparably harm, and eventually destroy, the Iraq marshlands. Severe deforestation from overgrazing upstream, combined with more than a decade of drought in the Middle East, exacerbated the environmental problems to the point at which Minister Latif believes the marshes would soon have been history "even without Saddam."

Dream boat. Iraqi expatriate Azzam Alwash envisions Marsh Arabs and ecotourists taking advantage of fully restored wetlands.

Brutal ecocide

But it was Saddam Hussein's regime that delivered the coup de grâce. Part of the Iran-Iraq border runs through the wetlands, and during the 1980s war, both sides built causeways and drained marshy areas for better access to the front. After the first Gulf War and the unsuccessful uprising of Shiite Muslims in the south, the Iraqi government set about draining the remaining marshes. Its goal was to remove the threat of insurgency and replace the marsh culture of fishing and rice production with dry agriculture. Massive dikes and canals were built to divert water from the marshes, quickly turning them to desert.

"The demise of these once-vast wetlands has been hastened through deliberate drainage by the Iraq regime," the U.N. study notes. To Latif, Saddam's actions constitute "a brutal ecocide" as well as "a crime against humanity." Many of the marsh inhabitants are now returning home, although most remain scattered in Iranian refugee camps or in cities.

The marshes are actually three distinct regions, each with its own particular ecosystem. The once vast Central Marsh, which covered more than 3000 square kilometers in 1973, has shrunk by 97%. Most of what remains are reeds growing in irrigation canals. Another marsh, called al-Hammar, lost 94% of its area, and al-Hawizeh, which borders Iran, is two-thirds smaller than 3 decades ago. Even the Hawr al-Azim Marsh, which is the Iranian extension of al-Hawizeh, is less than half its size due to reductions in water flow from Iraq.

Their depletion has led to the extinction of an otter, bandicoot rat, and a long-fingered bat particular to the marshes, and 66 species of water birds are at risk. Aquatic animals also have suffered, from shrimp to fish, with devastating consequences for coastal fisheries. "The wetlands were like a vast sewage treatment plant for the Euphrates and Tigris system," says Hassan Partow, who helped write the U.N. report. "They were the kidneys." Without them, the patient is imperiled.

Just add water?

How those kidneys function is uncertain. For decades, foreign and even Iraqi researchers were forbidden to enter the marshes, and in the 1990s the government destroyed a research station in the Hammar. As a result, most studies have relied upon Landsat remote-sensing data.

After the U.S. invasion in 2003, foreign scientists suddenly gained access. As they scramble to create an extensive database, it's hard to stay ahead of the population. When Saddam's regime collapsed, local residents jubilantly broke open the dikes and dams, reflooding nearly half of the marshes. "They did not wait for us," says Alwash.

Ancient battleground. Relief from the palace of Assyrian King Sennacherib, who sent troops to ferret out rebels in the species-rich Mesopotamian marshes in the late 7th century B.C.E.

The reflooding has been haphazard, however, and many dikes and dams from the Saddam era remain. But there is now plenty of water available: The U.S. invasion coincided with the end of a long drought in the region. Given the fast pace of dam construction in countries upstream and the possibility of another drought, though, renewed desertification is likely.

Last April a team funded by the Italian government began the first in situ study of the marshes, focusing on a small marsh of about 200 square kilometers called Abu Zirig. Reflooded in 2003 by locals, it has recovered rapidly. "This is a happy example," says Italian hydrologist Andrea Cattarossi. "The environmental conditions were pretty good." Carp, trout, smaller fish, and nearly half of the 50 to 60 species of birds that once flourished in the marsh have returned.

Although the marsh appears to be recovering, Cattarossi's data show that the volume of water is preventing light from reaching the roots of aquatic plants, threatening their growth. He says new structures are needed to control water circulation. He also worries about overfishing by a culture grown accustomed to using pesticides, explosives, and electrocution. A drought year will spell doom for Abu Zirig, he warns, because the water source first flows through agricultural areas.

But Abu Zirig is located north of the three main marshes and therefore receives larger quantities of less saline water than wetlands downstream. Those areas to the south will be more difficult to restore, scientists say.

CREDITS (TOP TO BOTTOM): DANA SMILLIE; WERNER FORMAN/CORBIS

Published by AAAS

Curt Richardson, an ecologist at Duke University in Durham, North Carolina, and lead author on the *Science* paper, used U.S. Agency for International Development (USAID) funding to examine other marsh regions during two visits. He says that reflooded portions of the eastern Hammar marsh are essentially saltwater deserts. "We're not quite sure why," he says. But he found that the area suffers from high salinity and high levels of hydrogen sulfide, which inhibits plant growth. Located close to the Persian Gulf—at the terminus of the "kidneys"—the marsh must be flushed with clean water to remove the salt and hydrogen sulfide. "The real question is whether there is enough water to do so," says Richardson.

One silver lining to the grim security situation, in which foreigners are often targets, is that Iraqi researchers are taking the lead in gathering data. "We had to train these people," says Cattarossi—no simple task given the lack of equipment and expertise following nearly 2 decades of Iraqi isolation. "And the situation in the field is very difficult; the vegetation is thick, and the [residents] can be a little bit suspicious." Richardson says that Alwash helped build trust with locals to ensure a steady flow of data for his study.

More than two dozen Iraqi biologists now help gather data from the Hawizeh, Hammar, and Central marshes for foreign researchers such as Richardson. Iraqi scientists declined to be interviewed, fearing reprisals from insurgents. But Ali Farhan, an Iraqi engineer and adviser to the Iraqi government, explains that two teams visit each marsh every month. They gather data on water quality, phyto- and zooplankton, bottom sediments, fish, and birds. Such practical training, he says, should help the shattered Iraqi scientific establishment gain a place at the table in the marsh discussions. So far, he adds, data gathering has gone smoothly, thanks to careful cultivation of local sheiks. But he says researchers are still harassed at times by the bandits who roam the region.

Mud to dust. The lush marshes around Abu Subat, unproductive desert after Saddam Hussein diverted water in the 1990s (*top*), are now trying to make a comeback.

CREDITS: AZZAM ALWASH

Romance or realism

During the next year, scientists using Italian funding hope to map the current water flow in the Iraq marshes as a first step toward understanding how to stabilize and revitalize the marshes. The U.S. Army Corps of Engineers is working up a model that details the flow of the Tigris and Euphrates. Meanwhile, the Center for Restoration of the Iraqi Marshes (CRIM), an organization of several Iraqi ministries created last fall in Venice, will put together a "master plan." "We need an international and Iraqi consensus," says Thomas Rhodes, an American ecologist and head of USAID's southern Iraq region based in Basra.

So far that's been elusive. Marsh advocates such as Alwash yearn for a vast reclamation project, with ecotourism to fuel the anemic local economy. He says it could be done for as little as $100 million, using Iraqi labor. Alwash worries that USAID, which is focusing on replanting date palms and ensuring the health of farm animals in southern Iraq, fails to grasp the enormous agricultural and economic benefits to the Marsh Arabs of a successful restoration.

But others dismiss such visions as unrealistic, given the current fiscal, social, and security situation. USAID has recently discontinued its funding of Alwash's organization, and Andrew Natsios, head of the agency, says, "the marsh people are not interested in restoring all the marshes. They want some restored and some left dry for agriculture."

Peter Reiff, an anthropologist who works on the marsh issue as a USAID contractor, notes that local Marsh Arabs have abandoned their old way of life cultivating rice and using water buffalo: "They are becoming farmers, and they get better returns with sheep, wheat, and cattle." And he accuses outsiders such as Alwash of possessing "a wistfulness about the marshes that is almost romantic. You are not going to make this into an ethnographic museum."

Iraqis themselves are divided. Minister Latif says he is "not very keen on the word "restoration"—restoring does not help the population." He favors a plan that focuses on health, education, and transportation needs. And the long-suffering local people appear to want it all. During two recent conferences on the marshes held in southern Iraq, participants urged that the wetlands be restored—and that new schools, clinics, and roads be built to lift them out of their dire poverty.

CRIM is supposed to bring all these disparate parties together. But some doubt it has the political and financial muscle to do so. "CRIM is understaffed, under-resourced, undertrained, if well-intentioned," says one foreign scientist involved in its creation. And negotiating a deal with Turkey, Syria, and Iran on water rights—a crucial element in any restoration plan—poses a daunting diplomatic challenge. "It is quite obvious that there isn't enough water to restore all the desiccated marshes," says Farhan.

Two cultures

There is consensus on the need to preserve at least some of the marsh. "There is great potential to restore a portion," says Cattarossi. Rhodes sees a scattering of core areas of protected, healthy, and biodiverse marsh surrounded by zones of compatible use. The cores would support ecotourism, provide a haven for animal and bird life, and allow the old marsh life to survive and even flourish alongside a more typical lifestyle of dry farming. "There will be enough room for both cultures," says Farhan. "In fact, Marsh Arabs are currently practicing both cultures."

To sell that vision, advocates argue that marsh restoration offers more than environmental and social benefits. "If we can get some of the marshes back, it would [also increase] security and stability," says Barry Warner, a biologist at the University of Waterloo, Canada, who is organizing a bird count this year in the marshes. Adds Natsios: "U.S. support for the marsh people is also protection for our troops."

Making that case will be tough, however. Last year, the U.S. Congress rejected spending any money on marsh restoration. But Natsios says his agency will continue to provide modest sums for marsh research and planning.

Even the enthusiastic Alwash—who calls the Eden Again project "his mistress"—acknowledges that full restoration is unrealistic, given the constraints of water, money, and political will. But he has not abandoned his dream of threading his way through endless beds of high reeds. Iraqis and their foreign friends may not be able to reconstruct a Garden of Eden. But they are hoping for the chance to recreate at least a piece of paradise.

–ANDREW LAWLER

Published by AAAS

WASHINGTON, D.C.—Befitting its location in the nation's capital, this year's meeting of AAAS (publisher of *Science*) from 17–21 February addressed the nexus of science and society. More than 9000 attendees heard about great strides in robotics (*Science*, 18 February 2005, p. 1082), celebrated the Year of Physics and Einstein's legacy (*Science*, 11 February 2005, p. 865), and presented research that spanned science, medicine, and politics.

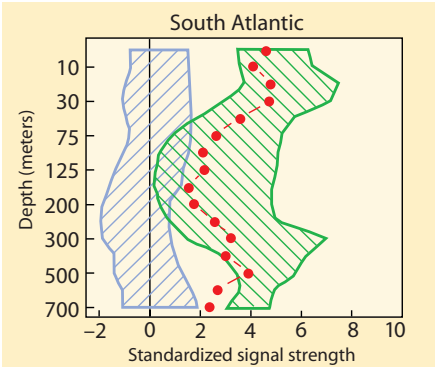
Ocean Warming Model Again Points to a Human Touch

Climate researchers concerned that their model might have overlooked something have retested the links between the burning of fossil fuels, greenhouse warming, and the warming of the deep oceans. A closer look at the evidence, they say, has bolstered their earlier conclusion: Humans are indeed warming the world, right down to thousands of meters deep in the oceans.

ONLINE COVERAGE
For additional stories from the AAAS meeting, see *Science*NOW (sciencenow.sciencemag.org)

The high statistical significance of the new study reported at the AAAS meeting "should wipe out much of the uncertainty about the reality of global warming," says the study's lead author, climate researcher Tim P. Barnett of Scripps Institution of Oceanography in La Jolla, California.

For a 2001 study (*Science*, 13 April 2001, p. 270), Barnett and colleagues ran a state-of-the-art climate model that traced where and when atmospheric heat trapped by rising greenhouse gases of the past century would have entered the oceans. When they compared the model's simulation to actual measurements of ocean temperature, they found a good match. With a confidence of 95%, they calculated, human-produced greenhouse gases are behind real-world warming. Three additional studies using three other models have yielded similar results.

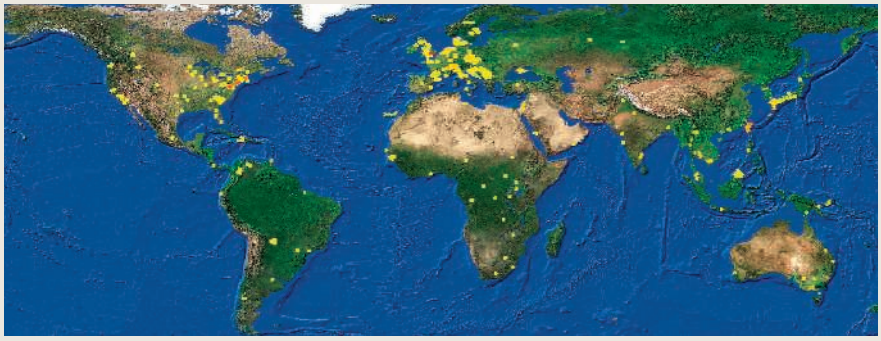


A match. A model's warming (green) fits the actual warming (red) and exceeds climate noise (blue).

Barnett and colleagues at Scripps and Lawrence Livermore National Laboratory in California, have now checked the model against better data, paying more attention to possible uncertainties in the model. They used a revised and updated set of ocean observations, this time avoiding a quirk in the data processing that had skewed temperatures in the data-poor Southern Hemisphere. To take account of variations among models, they compared detailed results from a second, independent model and studied how ocean warming in eight other models would affect the results. Unlike most other studies, they also followed the heat deep into separate ocean basins. In the end, the two main models "absolutely nailed the greenhouse signal" seen in the ocean, Barnett says. This time, statistical confidence is much greater than 95%, he says.

Most climate scientists are reassured. "The fact that multiple models simulate a comparable [ocean] warming gives a robustness to the results," says climate modeler Thomas Delworth of the Geophysical Fluid Dynamics Laboratory in Princeton, New Jersey. But climate researcher and modeler Gerald North of Texas A&M University in College Station still wonders whether the models have realistic enough oceans. More tests no doubt await.

–RICHARD A. KERR



More Infectious Diseases Emerge in North

A preview of a new global map of emerging infectious diseases turns a common assumption on its head. The map, presented in D.C. by Peter Daszak of the Consortium for Conservation Medicine at Wildlife Trust in New York City, spans the years 1940 to 2004 and indicates roughly 500 locations around the world where specific diseases first emerged. (Red indicates multiple events.) The map suggests that the majority of emerging diseases originated in Europe, North America, and Japan—a result that appears to hold up after correcting for reporting biases, according to Daszak and his co-workers. The media and funding organizations tend to assume that most infectious diseases emerge in the tropics because AIDS, severe acute respiratory syndrome, Ebola, and other high-profile diseases began there, says Daszak. But the preliminary map suggests that food-borne infections and drug-resistant microbes in the northern industrialized countries—the result of factors such as agricultural practices, the overuse of antibiotics, and international travel—are a more significant public health threat. "It's very counterintuitive to what most people think about emerging diseases," says Joshua Rosenthal of the Fogarty International Center at the National Institutes of Health.

–JOCELYN KAISER

Whaling Endangers More Than Whales

For more than a decade, scientists have gone to great depths to study the unusual deep-ocean communities known as whale falls. When a whale dies, its carcass sinks to the sea floor and provides a long-lived home to worms, clams, mussels, and many other creatures. New results presented at the AAAS meeting suggest that commercial whaling, even at so-called sustainable levels, would drive many of the novel species found at these cetacean gravesites to extinction.

In 1987, while surveying the sea floor in the submersible *Alvin*, biological oceanographer Craig Smith, now at the University of Hawaii, Manoa, and co-workers came

CREDITS (TOP TO BOTTOM): T. P. BARNETT; PETER DASZAK *ET AL.*

Published by AAAS

Homeless? Whaling threatens the survival of creatures such as this worm that live off whale carcasses.

CREDITS (TOP TO BOTTOM): ADRIAN GLOVER; CHRIS MADDALONI

across a strange menagerie thriving on and near a submerged whale skeleton. Since then, he and other researchers have shown that dead whales, like hydrothermal vents and cold seeps, can for decades support their own deep-sea biological communities. After all, the carcass of a great whale deposits up to 160 tons of blubber, meat, and bone in one fell swoop.

To estimate how 2 centuries of commercial whaling has affected whale-fall communities, Smith and his colleagues have now combined whale-population estimates with two ecological models: one that links habitat loss to biodiversity and a second that estimates how abundant a species must be to avoid going extinct. The best published estimates indicate that 75% of whale populations—and therefore whale-fall habitats—have been lost in the North Atlantic since large-scale whaling began in the early 1800s, Smith says. Based on those numbers, both models predict that whaling has already caused about 40% of North Atlantic whale-fall species to go extinct. Even at the so-called sustainable levels of whaling being considered by the International Whaling Commission, in which whale populations would be maintained at 50% of historic, prewhaling levels, 15% of the whale-fall species will disappear forever, according to the models.

In an ambitious effort that may help identify whale-fall species before they do go extinct, Smith and his colleagues have recently towed out to sea the huge carcasses of five whales that had beached themselves and died, sunk them, and periodically returned to each carcass in a submersible. Sleeper sharks, crabs, and hundreds of hagfish munched away at the carcasses for months, and many thousands of pea-sized amphipods nibbled on the smaller pieces. A strange assortment of creatures then colonized the bones and nearby nutrient-rich sediments, including several new species of the bone-eating zombie worm (*Science*, 30 July 2004, p. 668), which uses symbiotic bacteria to help digest the fatty marrow of whale bones. Over many decades, these microbes and other free-living bacteria break down oils trapped in whale bones, producing sulfide that fuels the growth of an average of 185 species per large whale skeleton. From these studies, the Hawaii team has upped the tally of species potentially unique to whale falls to 32.

"It's absolutely fascinating," says biological oceanographer Steven Palumbi of Stanford University in California. The work, he adds, shows that "there's a whole community of organisms in the deep sea that specializes in being the undertaker of these whale carcasses." And, says Palumbi, the modeling demonstrates that human activity can drive at least some deep-ocean creatures, besides whales, to extinction.

—DAN FERBER

DNA Tells Story of Heart Drug Failure

Physicians have long known that a drug commonly prescribed for heart failure helps only about half the patients who receive it. Now, researchers are explaining that puzzle using genetics. A study reported at the AAAS meeting found that a difference of a single amino acid within the drug's protein target may determine whether the drug works. The discovery could ultimately help physicians better juggle drugs in heart failure patients and possibly in those with high blood pressure as well.

In the late 1990s, pulmonologist Stephen Liggett of the University of Cincinnati, Ohio, along with his colleagues, found that people had a certain polymorphism, or genetic variation, in the gene encoding the beta-1 adrenergic receptor. That's the receptor targeted by the heart drugs known as beta-blockers.

In the general population, there are two common forms of the receptor gene: One version makes the receptor with arginine at a particular site; the other, which varies by just a single nucleotide, places a glycine there instead. Because every person has two copies of the receptor gene, inheriting one from each parent, an individual can have two copies of the glycine variant, two of the arginine, or one of each. Mice endowed with the human arginine variant are both more susceptible to heart failure and more responsive to beta-blockers, raising the possibility that the same holds true in people.

So, Liggett's team recruited 1040 volunteers, all people with severe heart failure. Roughly 490 had two copies of arginine, 450 had one of arginine and one of glycine, and the rest had two copies of the glycine version. Patients were randomly assigned to receive either a placebo or the drug bucindolol, a beta-blocker.

The researchers found that the cohort with two copies of the arginine variant were helped most by the drug. Compared to the placebo group, the bucindolol users experienced fewer deaths and hospitalizations over about 2 years (and in some cases up to five). Over the course of the study, 82% of them survived compared to 65% of those on the placebo. But those with one copy of each receptor variant or two copies of the glycine version weren't helped at all by the drug, faring about as poorly as the patients on placebo who had two copies of the arginine variant.

Liggett intends to put together a larger study to confirm the findings. This time, however, all patients would receive the bucindolol because it wouldn't be ethical to give double-copy arginine patients a placebo, Liggett said during his presentation.

QUOTE

"I don't see any evidence of a backlash by this Administration [against supporters of Democrat John Kerry]. If we were to withhold funding from every scientist who was a Democrat, there wouldn't be much science."

—Presidential science adviser John Marburger, discussing the politicization of science at a pre–AAAS meeting workshop sponsored by the National Association of Science Writers.

Liggett notes that, on its own, the variation in the beta-adrenergic receptor gene doesn't seem to affect heart failure risk. But people with two copies of the gene for the arginine variant and two copies of another gene—a combination nearly unique to African Americans—have 10 times the risk of heart disease.

"He's found a polymorphism that seems to predict response" to bucindolol, says molecular pharmacologist Kathy Giacomini of the University of California, San Francisco. The result, she adds, is "very exciting" and "very specific." It's not clear yet, says Liggett, whether the findings apply to other beta-blockers, which are also used to treat high blood pressure.

—JENNIFER COUZIN

Ancient Alexandria Emerges, By Land and By Sea

Excavators are finding surprisingly late signs of intellectual life in the ancient capital of Hellenistic Egypt and discovering that geology played a dramatic role in the city's fall

OXFORD, U.K.—For centuries the massive Pharos lighthouse, one of the seven wonders of the ancient world, guided sailors to the busy wharves that made Alexandria a prosperous center of Mediterranean culture and home to the greatest library of ancient times. Yet while rivals Rome and Constantinople survived the chaotic period following the collapse of the Roman Empire, Alexandria faded from the historical record. By the 8th century C.E. the famed metropolis had fallen into oblivion.

An ancient wonder. Excavators seek the remains of the Pharos lighthouse, shown here in a Renaissance artist's view.

Today the city of Alexandria, site of Alexander the Great's tomb and Cleopatra's death, attracts scholars the way it once drew merchants and philosophers, as shown by a recent conference at Oxford University.* Rescue archaeology amid rapid urban growth combined with new underwater mapping technologies are yielding new insight into the old city's role and history. Archaeologists have uncovered tantalizing hints of surprisingly early beginnings as well as signs that the city's vibrant intellectual life lasted far longer than anyone had expected. "Now we can imagine the functioning of a university in antiquity," says historian Manfred Clauss of Germany's University of Frankfurt.

New data also suggest that environmental disaster played an important role in ancient Alexandria's downfall, which has long been attributed primarily to religious and political turmoil; the fate of Alexandria could provide a warning for today's fast-growing cities built on deltas, researchers say.

* "City and Harbour: The Archaeology of Ancient Alexandria" at the Oxford Centre for Maritime Archaeology, 18–19 December 2004.

An ancient think tank

Most of the new data comes from digs that began in the 1990s, when the Egyptian government lifted a ban on underwater archaeology and began to encourage salvage work on land as the city of 6 million expanded over its ancient foundations. Those foundations were officially laid in 332 B.C.E., when legend has it that the Greek bard Homer appeared to Alexander the Great in a dream and urged him to found a city along the narrow strip of land separating Lake Mareotis from the Mediterranean Sea. It seemed an unlikely choice; the mouth of the Nile was far to the east, where the major Egyptian ports were well established, and only small villages existed on the spot.

The city grew to prominence after Alexander died in 323 B.C.E., and his general Ptolemy made it the capital and largest port of Egypt. The vast wealth of Ptolemy and his Greek and Macedonian successors built a Greek-style city of temples, lavish palaces, and the famous library, says historian Gunther Grimm of Germany's University of Trier; scholarship in philosophy, physics, mathematics, and astronomy thrived. Under the Roman rule that followed Cleopatra's death in 30 B.C.E., Alexandria served as the nexus for grain exports for the vast empire. But within a few centuries, the city largely vanishes from the historical record.

Now nearly a dozen teams of excavators are sifting through what remains both in the city and in the harbor. Archaeologist Jean-Yves Empereur of the Center for Scientific Research in Paris is working fast to salvage remains of the ancient city as the new one expands. "Ten to 12 meters beneath the modern city, the old one is very well-preserved," says Empereur, who was one of

Oxford Center Raises Controversy

Ancient Alexandria was famed for its philosophical disputes, and that tradition is very much alive in excavations now under way in the Egyptian port. Scholars are hotly debating a controversial agreement that gives a nonscientist, French businessman Franck Goddio, control over underwater archaeological data collection for Oxford University. At a conference held in December—a coming-out party for Oxford's new Center for Maritime Archaeology—dozens of scholars discussed new finds (see main text). But others avoided the event, arguing that contracting out the leadership of maritime digs to nonscientists sets a poor precedent.

Under the deal signed 18 months ago, Goddio will oversee undersea excavations; Oxford graduate students, under the guidance of professors, will analyze the data. The Hilti Foundation of Lichtenstein, which has supported Goddio's work for a decade and is funded by a tool company of the same name, will provide at least $300,000 to fuel the center, which for now will focus on Goddio's work in Alexandria and nearby Abukir Bay.

Goddio, who has worked for 20 years on more than 50 underwater sites, is a Jacques Cousteau of archaeology, often featured on European television. Although he lacks a degree in archaeology—he studied statistics—Goddio says his experience speaks for itself. But he also seeks academic respectability. The Oxford agreement "is a chance for us to get closer to a university which could back our work and take advantage of our discoveries," says Goddio. "We were looking for a scientific base or 'harbor' for the findings and results from Franck Goddio's excavations," adds Michael Hilti, who heads the Hilti Foundation. "With Oxford, I think we have found a perfect partner."

The arrangement makes sense to university officials, who are eager to enter the burgeoning and expensive field of maritime archaeology. "We were blown over by the quality of Franck's underwater fieldwork," recalls Barry Cunliffe, the Oxford classical archaeologist who helped broker the deal. "It was an extremely smart piece of archaeology, well-ordered and observed."

But Goddio's deal with Oxford raises concerns among many maritime archaeologists uncomfortable with turning over part of the scientific process to those who lack formal training. "I'd be wary of

CREDIT: BETTMANN/CORBIS

Published by AAAS

several scholars who declined to attend the conference because of their concerns about Oxford's ties to a private underwater archaeologist, Franck Goddio (see sidebar). Empereur adds that the houses, streets, and mosaics that have been uncovered represent "just 1% of what could be rescued."

Even that small percentage is rewriting the city's history. Most historians assume that intellectual life in the city withered with the destruction of the library—which likely occurred over hundreds of years—and the rise of Christianity. But among the most intriguing recent finds is a complex of lecture halls that appear to be "the center of [the city's] intellectual and social life in late antiquity," says Warsaw University's Grzegorz Majcherek of the Polish-Egyptian Archaeology Mission. Each hall includes a single central seat for a notable—likely the teacher—and often a smaller seat on the floor, perhaps for student recitations. The complex is part of the old city's most extensive area of urban architecture. Majcherek estimates that the halls were built in the late 5th and early 6th centuries C.E. and notes that a Roman theater was even converted into a lecture hall at this time. He speculates that what he calls "the Oxford of antiquity" could have survived into the era of Arab control—"surprisingly late."

The find intrigues historians, who say there has been little evidence that intellectual life in the city flourished for so long. "This is the most exciting find in years in Alexandria," says Clauss. "The buildings Professor Majcherek has found demonstrate the existence of a think tank" long after the fall of Rome. "It is surprising that it seems to function in a modern way," he adds.

Facing the past. Divers have found artifacts such as this sphinx representing Cleopatra's father.

Down under

Just a few hundred meters away, an important part of ancient Alexandria lies undisturbed underwater, meters from the modern breakwater lining the harbor. In the 1990s, Empereur uncovered statuary and blocks that may be portions of the Pharos lighthouse, which survived in ruins until an earthquake in the 14th century. Goddio found a sunken palace from the Ptolemaic era and brought up statues and other artifacts that he hailed as remnants of Cleopatra's palace. That claim, as well as the exact location of the Pharos, remains in dispute.

More recent finds are less spectacular, but they shed important light on the evolution of the harbor that was Alexandria's heart. For example, Goddio's team now has found evidence of a dock that dates to about 400 B.C.E., predating Alexander. "We were surprised, took new samples, and got the same answer—this was most probably a pre-Ptolemaic structure [and is now] 7.5 m below sea level," says Goddio. Geologist and team member Jean-Daniel Stanley of Washington, D.C.'s Smithsonian Institution told meeting participants he has found tantalizing hints that inhabitants smelted lead on the site as early as 2000 B.C.E.

The discoveries are part of an ambitious effort by Goddio to map the entire harbor bottom—one data point for every 25 centimeters—and conduct extensive radiocarbon dating of planks and pilings brought up by divers. The survey of the 2.5-kilometer-by-15-kilometer area will give researchers "quite a precise idea" of the location of docks and buildings that lined the harbor, says Goddio: "A ghost from the past is being brought back to life."

Meanwhile, geologist Stanley has examined dozens of cores from the harbor and

Modern mariner. Former businessman Franck Goddio leads spectacular underwater digs.

CREDITS (TOP TO BOTTOM): FRANCK GODDIO/HILTI FOUNDATION/DCI; CHRISTOPH GERIGK/FRANCK GODDIO/HILTI FOUNDATION

parceling out one of the links in the chain of science," says maritime archaeologist Jon Adams of the University of Southampton, U.K. "Archaeology should be conducted by proven and trained archaeologists," adds George Bass, a professor emeritus at Texas A&M University in College Station who is considered one of the founders of maritime archaeology.

Robert Grenier, head of Ottawa's Parks Canada maritime archaeology unit, adds that Goddio's record is big on coffee-table books but small on scholarly publications. For example, he says, Goddio excavated the 35-meter-long Spanish galleon *San Diego* off the coast of the Philippines and produced a glossy catalog but limited scientific data. Grenier worries about data that may not be collected, such as apparently inconsequential fragments that might provide a clue to a ship's identity or place of construction.

Goddio defends his record, noting that the second *San Diego* mission lasted more than 4 months and was devoted to understanding the ship's hull construction; he adds that he still hopes to publish more details.

Cunliffe insists that skilled nonscientists can make an enormous contribution because retrieving information from underwater digs is so technologically intensive and expensive. The choice he sees is to ignore nonscientists' expertise and funding, or to find a creative way to work with it. "The cost of doing this work is almost prohibitive unless you have the backing of a large foundation," he says.

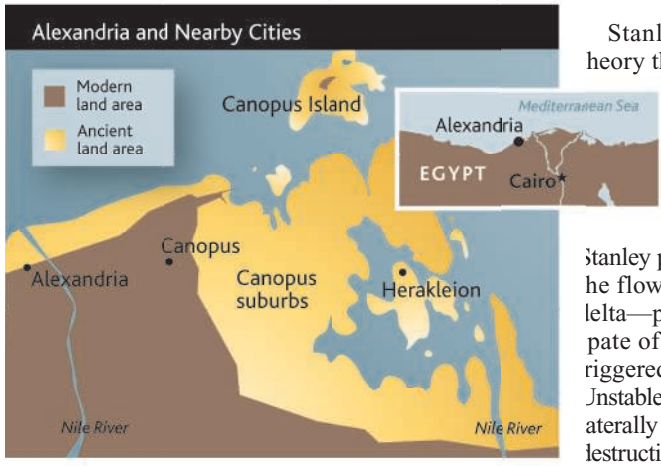
A large maritime excavation can cost upward of $1 million a month, forcing many underwater archaeologists to seek foundations or television producers to help fund their work. "We've all done a bit of whoring to get the money we need," admits one respected maritime archaeologist. And when it comes to Goddio's bountiful financial support, he adds, "I'm jealous." **–A.L.**

Published by AAAS

uncovered evidence of the centuries-long battle that ancient engineers waged against both gradual and sudden subsidence. He says the subsidence was brought on by a lethal combination of earthquakes, tsunamis, and the slow but relentless sinking of heavy foundations into unstable soil, which defeated even savvy Roman engineers. Although several wharves appear to have been reconstructed over centuries, no amount of piling could long hold up heavy stone foundations and buildings, he says. "[Adding] on all that material was asking for trouble," Stanley says. "The additional weight of a wave surge could be powerful enough" to submerge part of Alexandria's shore.

The historical record also shows an unusually active period of tremors from the 4th to the 6th centuries C.E. Quakes and tsunamis could have transformed sediment into a more fluid state, says Stanley. Sixty-five cores taken from the western harbor show signs of ancient liquefaction, he said, and numerous pieces of red coral not native to the harbor suggest that a tsunami washed them into the basin. But he says it is too early to reconstruct details of ancient collapse and rebuilding. "We need better 3D images of harbor substrate" to understand what repairs were done and when, he said.

The impact of these geological forces extended beyond Alexandria—and with even more dramatic consequences. Stanley and Goddio also are excavating three submerged cities in nearby Aboukir Bay: Herakleion, Canopus, and Menouthis. The first was an important entrance point to the mouth of the Nile, and the others were well-known pilgrimage sites. The area received huge amounts of sediment from the Nile, which compacted and sank over time. This process, combined with a slow rise in world sea levels, pushed the water at least 2 meters higher between the 6th century B.C.E. and the 7th century C.E. "Arabic texts show a huge Nile flood in 741 and 742 A.D.," notes Clauss. And by the 8th century—the same time Alexandria slips into obscurity—the historical record on these sites goes silent.



Stanley's research supports a theory that combines catastrophe and gradual sinking to explain the disappearance. Submergence alone cannot explain why much of the area is now a full 6 meters under water, and Stanley posits that sudden shifts in the flow of Nile branches on the delta—perhaps brought on by the spate of earthquakes—may have triggered more dramatic changes. Unstable sediment would have been laterally displaced, causing sudden destruction as the Nile moved into a new bed. Goddio's team has found evidence of human remains underneath toppled walls at the three sunken cities, backing up this theory.

In an era of climate change and fast urban growth along coasts, this research may have implications today. Stanley notes that modern cities such as Venice, Bangkok, and New Orleans sit on unstable delta soils. "This is becoming a world problem," he says. "Understanding the subsidence threat might help such cities avoid the fate of Herakleion, Canopus, and Menouthis." Alexandria, at least, escaped with only a flooded harbor—a sign that Homer perhaps was as canny a geologist as he was a storyteller. **–ANDREW LAWLER**

Radio Astronomy

Bristling With Promise

By substituting software for massive signal-gathering dishes, arrays of simple FM antennas offer astronomers a cheap, versatile alternative to traditional radio telescopes

In a remote Chinese valley sit 25 neat clusters of antennas, each tipped slightly askew. They are testing the airwaves, listening for interference from TV signals. If reception is clear enough and other things go well, within the next year or two the fields of the Ulastai Valley will fill with tilted antennas, like a Christmas tree farm pummeled by wind.

The valley will become a huge array of 2-meter-long antennas, 10,000 strong, covering 30,000 square meters. The array, dubbed the Primeval Structure Telescope (PaST), is the brainchild of a group of Canadian, Chinese, and American scientists pursuing a low-frequency portrait of the early universe. And they hope to find it out in the vast, quiet stretches of western China, one of the last places on Earth out of the reach of jabbering TV and FM radio broadcasts.

Though just 25 pods of 127 antennas each right now, PaST is a herald of what's to come. Thanks to recent advances in theory and computing power, radio astronomers can now build telescopes consisting of huge arrays of antennas capable of viewing the universe in a novel palette of low frequencies hitherto rarely used for astronomical observations. "What's most exciting to me [is] that we don't know what we're going to see," says PaST collaborator Jeffrey Peterson of Carnegie Mellon University in Pittsburgh, Pennsylvania.

Peterson isn't alone in his enthusiasm. Several other array telescope projects are under way in the Netherlands, Western Australia, and the American Southwest. Their scientific goals include finding radio equivalents of gamma ray bursts and detecting the faint traces of the first stars.

The arrays will take radio astronomy back to its roots in the 1930s, when Karl Jansky, an engineer at Bell Telephone Laboratories in Holmdel, New Jersey, noticed radio waves emanating from the center of the Milky Way galaxy at 20 MHz. The field took off after radar operators during World War II discovered a technique called interferometry, which enabled astronomers to string together several small antennas to get the same resolution as that of one huge antenna. But researchers soon realized that low frequencies were wrinkled and warped into indecipherability by Earth's ionosphere. Frustrated, they switched their attention toward frequencies above 1 GHz. And that's where radio astronomy stayed until recently, when new calibration techniques opened a window into the low-frequency range.

The breakthrough came in 1991, when astronomers using computer algorithms to correct for the effects of ionospheric interference jiggered the Very Large Array (VLA), a Y-shaped assemblage of 27 dish antennas in western New Mexico, into receiving at a record low frequency of 74 MHz. Their success blasted open opportunities for large low-frequency arrays. "The old low-frequency telescopes were like a nearsighted person trying to read from far away without his glasses," says Namir Kassim, a radio astronomer at the Naval Research Laboratory in Washington,

SOURCE: ADAPTED FROM FRANCK GODDIO/HILTI FOUNDATION

Published by AAAS

D.C. "When we learned how to put glasses on these telescopes, all hell broke loose, and now everyone's trying to build them."

VLA and its kin, however, suffer from a major disadvantage: Dish antennas are not good at picking up low frequencies. If the length of the wave is close to the diameter of the dish or longer, the dish can't see it at all. By contrast, arrays of wire antennas—either simple FM dipoles or log-periodic antennas made of multiple dipoles—can be designed to pick up any wavelength astronomers might fancy. All it takes is the right arrangement, enough land, and a supercomputer programmed to convert the jumbles of waves sweeping across the arrays into useful images. Somewhere between 200 MHz and 100 MHz, the efficiency advantage switches over from dishes to dipoles.

In the early 1990s, inspired by the new calibration techniques, scientists from the United States, the Netherlands, and Australia started investigating designs and locations for the Low Frequency Array. LOFAR was conceived as a cheap, quick instrument to get a first look at the low-frequency universe, says Kassim, who was LOFAR's international project scientist. But ballooning costs and disagreements over siting and scientific goals crippled the project, and the collaboration disintegrated.

Today, Germany and the Netherlands are working to build a 25,000-antenna LOFAR array spread over an area 350 kilometers in diameter on both sides of their border; researchers hope to develop sophisticated new techniques to filter out radio and television signals. Other erstwhile LOFAR partners, meanwhile, have hatched projects that rely on geography to solve the problem of interference. Kassim and colleagues at the Naval Research Lab are now collaborators in the Southwest Consortium, a low-frequency radio astronomy project based in the American Southwest, a popular site for traditional radio telescopes. The telescope, known as the Long Wavelength Array (LWA), will examine ultralow frequencies from 20 MHz to 80 MHz with 10,000 dipoles strung along 400 kilometers.

Other former LOFAR participants—astronomers at the Massachusetts Institute of Technology (MIT), the University of Melbourne, and the Australia National Telescope Facility—have joined with the Harvard-Smithsonian Center for Astrophysics to plan a 3000-antenna low-frequency array in the outback of western Australia. If all goes well, the Mileura telescope will look at large-scale structure in the universe. The United States and Australia are negotiating joint funding for the $10 million project.

Location is key. FM radio and television signals are as damaging to radio astronomy as light pollution is to optical astronomy and are far more pervasive. "The best place for this would be the far side of the moon," says Jacqueline Hewitt, an astronomer at MIT involved in the planning of the Mileura project. "But if we can't go to the moon, we'll go to western Australia." PaST scientists say that if they could afford to, they would build their array at the South Pole, 2000 km from the nearest TV transmitter and with a 6-month polar night to increase the transparency of the ionosphere.

Array of hope. Peterson predicts a bright future for PaST.

CREDIT: COURTESY OF J. PETERSON

Remote real estate aside, antenna arrays offer several advantages over traditional radio telescopes. For one, their cost—a fraction of the $3.5 million price tag of a single VLA dish—puts them within reach of even small astronomical partnerships. In the spring of 2003, for example, PaST was just an idea Carnegie Mellon's Peterson was batting around with Ue-Li Pen of the University of Toronto. Within a year, Xiang-Ping Wu of the National Astronomical Observatory of China had joined the project, and they had gotten $600,000 of funding from the Chinese government, found a site, and set up test antenna. The current funding will support an array of 2500 antennas. The researchers expect that support from the National Astronomical Observatory will enable them to install another 7500 antennas by 2006.

Antenna arrays boast technical advantages as well. Existing radio dish telescopes, such as Arecibo in Puerto Rico, just aren't up to the task of seeing in the ultralow frequencies, the researchers say. Big dish telescopes with just one, huge antenna have too narrow a field of view. They can sweep the sky over a period of time, but sweeping complicates the delicate calibration needed to compensate for the ionosphere. Dish arrays have the same drawback. Dipole arrays, by contrast, can effectively look in any direction or all directions at once, as long as computers are available to crunch the data. Log-periodic arrays are similar but have narrower fields of reception.

The 180-degree view of a dipole array means that transient phenomena may be noticeable. For example, researchers say, a very high-resolution array might pick up bursts of ultralow-frequency radio waves from gas-giant planets circling other stars—something a dish telescope could spot only if it were looking in exactly the right direction at the right time. Arrays might also pick up long-theorized radio counterparts to the cosmic energy blasts known as gamma ray bursts. By studying how such radio bursts distort as they cross space, astrophysicists could test cosmological models that predict there should be lots of ionized gas between galaxies.

The most mouthwatering possibility is that sensitive array telescopes will catch whispers of reionization, the moment early in cosmic history when the first stars flickered on. At that time the universe was filled with neutral hydrogen gas. Ultraviolet light from newly ignited stars blew electrons off hydrogen atoms, creating ions that don't radiate at certain wavelengths. The result, if radio astronomers can spot it, should be big patches of silence between 100 and 150 MHz amid an otherwise steady hum of neutral hydrogen. PaST, Mileura, and LOFAR all hope to detect reionization, although researchers acknowledge that it's a long shot.

At least astronomers won't have to wait long for the arrays themselves. PaST spotted galaxies at redshifts of 0.3 last year with just two protopods, and it is scheduled to start collecting data seriously this spring. The Southwest Consortium could have the core of the LWA up and running by 2008. The Mileura project hopes to erect its first test antennas by the end of the month. And more ambitious arrays may be in the works. From now on, when the universe rumbles in the ultralow, we'll be listening.

–KIM KRIEGER

Kim Krieger is a freelance science writer in Washington, D.C.

Published by AAAS

RANDOM SAMPLES

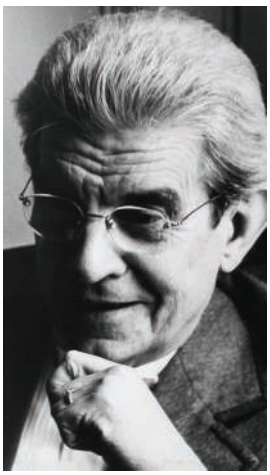
Edited by Constance Holden

French Psychoflap

Freudian psychoanalysis is far from the mainstream in modern mental health care. But it's alive and well in France—and it just got a shot in the arm from health minister Philippe Douste-Blazy, to the consternation of many scientists.

Speaking at a 5 February meeting of psychoanalysts in Paris, Douste-Blazy praised their work while announcing that he had ordered the removal from his department's Web site of a 2004 report concluding that the scientific evidence favors cognitive-behavioral therapy (CBT) over psychoanalysis. "You won't hear about [the report] again," Douste-Blazy, a cardiologist, assured his elated audience.

France has a strong psychoanalytical tradition, founded by Jacques Lacan (1901–81), who melded classic Freudian ideas with structuralism in what his detractors say is a pseudoscientific, cultlike movement now led by his son-in-law Jacques-Alain Miller. Many of its followers were angered when France's leading health agency INSERM issued a report in February 2004 that took the currently popular "evidence-based" approach to psychotherapy and concluded that CBT has the most to show for itself.



Post-Freudian Lacan.

This time, many other psychologists and psychiatrists are incensed. "I'm totally amazed and puzzled," says Jean Cottraux, a psychiatrist at the Pierre Wertheimer Neurological Hospital in Lyon and a member of the INSERM panel. He calls the report's removal "an act of censorship" that could favor a regressive "lacanist takeover" of the field.

There's speculation that Douste-Blazy's remarks also are behind the sudden resignation last week of epidemiologist William Dab, director-general for health, whose office had requested the study.

Sensitivity or Censorship?

A new wave of political sensitivity appears to be sweeping over the national parks, to the dismay of some archaeologists. This time it's focused on the word "Anasazi," a term used for almost a century by archaeologists to denote ancient pueblo dwellers of the Southwest.

At Colorado's Mesa Verde National Park, there are two books you can't find these days: *Understanding the Anasazi of Mesa Verde and Hovenweep* by David Grant Noble and *Water for the Anasazi* by Kenneth Wright. Some archaeologists such as David Breternitz, professor emeritus of the University of Colorado, Boulder, are upset at the exclusion, calling it "censorship." But Tessy Shirakawa, a park spokesperson and tribal liaison, explains that some Pueblo Indians consider the word "Anasazi" derogatory. The Navajo term can mean "Enemy Ancestors" or "Non-Navajo Ancestors," depending on the tribe, according to archaeologist Linda Cordell of the University of Colorado, Boulder, Museum. Many Indians therefore prefer the term "Ancient Puebloans." Shirakawa says the park is switching terms in its own publications and asks authors and publishers to do the same.

Ancient Puebloan spiritual undertaking.

There has also been a minor flurry over the matter of rock art at New Mexico's Petroglyph National Monument bookstore. The bookstore has declined to carry *Rock Art in New Mexico* by Polly Schaafsma. Some Indians have objected to the term "rock art" because "art" is said to suggest a European cultural activity rather than a spiritual undertaking, says Diane Souder, tribal liaison and monument spokesperson. However, Souder says the book was turned down for yet another reason: It shows sacred images that tribes believe should not be photographed. For this reason, Souder says, the bookstore no longer sells calendars with such images.

Math Without Words

How much higher thinking can there be without language? That's been a perennial question. An intriguing piece of the answer has been supplied by research with three brain-damaged men whose ability to use words is severely impaired but who can still do complex math.

Math and verbal reasoning clearly involve analogous functions, but experts differ on how the two types of syntactic processes are related. To cast more light on the question, a team led by Rosemary Varley of the University of Sheffield, U.K., administered mathematical tests to three men who had damage to the language-processing centers in their left cortical hemispheres. The tasks involved not only simple addition, subtraction, multiplication, and division but more complex problems such as numbers embedded in brackets analogous to language structures—subordinate clauses, for example—that the men were no longer able to use. Although they even had difficulty deciphering words for numbers, they performed competently on a variety of tests presented in Arabic numeral format, the authors report in a paper appearing online on 14 February in the *Proceedings of the National Academy of Sciences*.

The study offers "a powerful new argument" for the view that "in the adult brain at least, the syntax of arithmetic and the syntax of language are independent," says neuroscientist Brian Butterworth of University College London. The authors say the results don't rule out the possibility that language in early life provides a "template" that facilitates math learning. But by adulthood, it appears that mathematical thinking can stand on its own.

Published by AAAS

Edited by Yudhijit Bhattacharjee

JOBS

Departures at NCI. Two leaders of the $4.8 billion National Cancer Institute (NCI) have found greener pastures. J. Carl Barrett, a 28-year veteran of the National Institutes of Health who since 2000 had headed

NCI's Center for Cancer Research, is leaving to oversee the Novartis Institutes for Biomedical Research's global program in oncology biomarkers. And oncologist Karen Antman, who came to NCI from Columbia University last April to be deputy director for translational and clinical sciences, is leaving to head Boston University's medical school.

Barrett had been tapped to become NCI Director Andrew von Eschenbach's deputy for basic science, a position that von Eschenbach created when he arrived in 2002 but has yet to fill. Von Eschenbach said last week that although NCI is losing "two very key, critical members, … we are celebrating that loss" because they are moving to "positions of enormous importance and responsibility."

Physics all-star. The U.K. Institute of Physics has bagged a consummate all-rounder as its new chief executive. Robert Kirby-Harris not only has a degree in theoretical physics and a doctorate in higher education policy, but he has also taught high school, served in the Royal Navy, run a research and consultancy company, and been a senior administrator at universities in the U.K. and Namibia. His current job is corporate director of the Royal Botanic Gardens at Kew. Among other issues, Kirby-Harris must grapple with the decreasing popularity of physics among students (*Science*, 4 February, p. 668).

Meanwhile, one of the U.K.'s main sponsors of basic research, the Particle Physics and Astronomy Research Council, also has a new leader: space scientist

Keith Mason. The 54-year-old Mason is lead researcher on one of the telescopes for the Swift gamma ray mission, launched in November. Mason's 4-year term will cover milestones such as the 2007 opening of the Large Hadron Collider at CERN.

Renaissance man. The new director of the Swiss National Science Foundation (SNF) is a political scientist and an expert on Florence during the Renaissance.

Daniel Höchli, 42, taught at the University of Saint-Gall for 7 years before joining Switzerland's Federal Department of Justice and Police in 1996,

where he became chief of staff of the Federal Office of Police in 2001. His administrative experience and his "personal love of science" make Höchli a great choice to run the $380 million agency, says Werner Stauffacher, vice president of SNF's Council and chair of the search committee.

Got any tips for this page? E-mail people@aaas.org

AWARDS

Never too late. The White House has announced the winners of the nation's top honors in science and technology for 2003.

Eight researchers will receive the National Medal of Science: J. Michael Bishop, University of California, San Francisco; G. Brent Dalrymple, Oregon State University; Carl R. De Boor, University of Wisconsin, Madison; Riccardo Giacconi, Johns Hopkins University; R. Duncan Luce, University of California, Irvine; John M. Prausnitz, University of California, Berkeley; Solomon H. Snyder, Johns Hopkins University School of Medicine; and Charles Yanofsky, Stanford University.

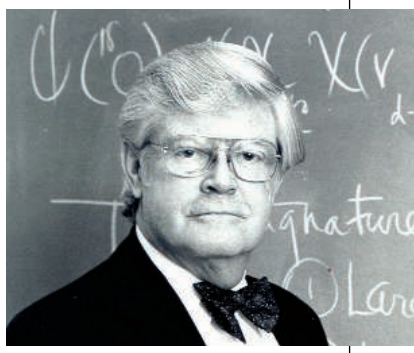
Six individuals, one company, and a foundation have won the National Medal of Technology: Jan D. Achenbach, Northwestern University; Watts S. Humphrey, Software Engineering Institute, Pittsburgh, Pennsylvania; Robert M. Metcalfe, Polaris Venture Partners, Waltham, Massachusetts; Rodney D. Bagley, Irwin Lachman, and Ronald M. Lewis, Corning Inc.; UOP LLC, Des Plaines, Illinois; and Wisconsin Alumni Research Foundation, Madison.

The awardees will be honored at a White House ceremony on 14 March.

DEATHS

Surely you're joking. D. Allan Bromley, who died of a heart attack on 10 February at the age of 79, had a distinguished career as the founder and longtime director of Yale's A. W. Wright Nuclear Structure Laboratory and science adviser to President George H. W. Bush. Less well known—but in keeping with his warm sense of humor—is his contribution to theological meteorology.

In 1972, Bromley was one of the brains behind an anonymous article in *Applied Optics* that combined clues from two biblical passages with physical principles to put bounds on the temperatures of heaven and



hell. Using the physics of radiation and the chemistry of sulfur, the authors argued that the bright light of eternal bliss suggested that heaven was a sweltering 798 K, whereas hell's boiling brimstone lakes had to be below 718 K. The paper quickly became physics legend and remains alive on the Internet.

In 1998, two Spanish physicists threw cold water on the analysis by pointing out an error in the paper's interpretation of biblical passages, bringing heaven's temperature down to a more tolerable but nonetheless scalding 505 K. When asked to comment on the new finding, Bromley laughed and said, "Go to hell."

CREDITS (TOP TO BOTTOM): SOURCE: INSTITUTE OF PHYSICS; BILL BRANSON/NCI; T. CHARLES ERICKSON/YALE UNIVERSITY; SNF

Published by AAAS

LETTERS

The Emergence of the ERC

AS SOMEONE WHO HAS BEEN AWARDED funding in multiple projects in both sides of the Atlantic, I see the emergence of an European Research Council (ERC) ("Decisive day for European research," W. Krull, H. Nowotny, Editorial, 5 Nov. 2004, p. 941) as an opportunity to improve on existing models. I would like to offer my two cents. First, it may not be that important that an ERC is well funded to the point that it could replace national granting agencies—recent reports suggest that this is most likely too tall an order. I see it as more critical that it should be funded and configured to provide unbiased supranational evaluation of research proposals. The rankings could then be used for awarding EU funding or could instead, or jointly, be carried by national agencies with a mandate to fund the best proposals in a given area. Second, U.S. granting agencies appear to be shifting from a paradigm that mostly targeted individual-initiated research to one that more likely rewards collective initiatives. In my experience, this actually better promotes young researchers and provides more protection to fundamental work under the umbrella of longer-term goals. This is particularly clear in the NIH roadmap. There may be a precious lesson here the ERC may want to consider: Micromanaging the progress of 3-year research programs falls significantly short of objectively evaluating the achievements of 5- or even 7-year larger initiatives.

JONAS S. ALMEIDA
Department of Biometry and Epidemiology, Medical University of South Carolina, 135 Rutledge Avenue, Charleston, SC 29425, USA.

Response

I AM GLAD TO SEE THAT THE ESTABLISHMENT of an ERC elicits comparisions and advice as offered by Almeida. It was never intended that an ERC would replace national granting agencies. Apart from being legally impossible, this would have meant political suicide and is indeed too tall an order to finance from the EU budget. But the proposal to use the unbiased supranational evaluation of research proposals, in case they are not funded, as a recommendation especially for national funding is an interesting one. The paradigm shift observed among U.S. granting agencies from targeting individual research to collective initiatives has long been preceded by EU funding practice, however. This has been the standard rule in all Framework Programmes since the beginning. In setting up only two criteria—attracting the best of the best in terms of scientific excellence, and a rigorous competition by supranational peer review—funding for basic research at EU level through an ERC should be open to individuals as well as to research groups across the European Research Area without any further strings attached. This is the truly innovative feature of an ERC—if and when it comes about.

HELGA NOWOTNY
Society in Science Program, ETH-Zentrum, CH-8092 Zurich, Switzerland.
*Chair of the European Research Advisory Board and member of the Expert Group examining the need for establishing a European Research Council

Protecting Privacy of Human Subjects

JOCELYN KAISER'S ARTICLE ON THE U.S. FEDERAL regulations known as the HIPAA Privacy Rules (*1*) describes these laws as overly complicated and obstructing life-saving research ("Privacy rule creates bottleneck for U.S. biomedical researchers," News Focus, 9 July 2004, p. 168). It ends with investigative pathologists calling for Congress to loosen the regulations to avoid lasting damage to research. In the same issue, Z. Lin *et al.* reach the opposite conclusion ("Genomic research and human subject privacy," Policy Forum, 9 July 2004, p. 183). Because they perceive the HIPAA Privacy Rules as not governing genetic information, they warn about privacy risks created by this oversight and recommend that these rules be strengthened—for example, by including strong penalties for privacy violations.

> **"[Legislation] would have to extend privacy protections to genetic information ... in an individually identifiable form."**
> —ROCHE

The problem with these articles is not that they present conflicting recommendations for regulatory policy, but that they include misinformation about the HIPAA regulations and how they govern research activities. The focal point of the HIPAA rules is health information received or created by health care providers (*2*), and when that information may be disclosed to others, such as researchers. Protected health information as defined in the regulations would include genetic information, and therefore, providers must abide by all the disclosure rules in regard to identifiable genetic information.

The HIPAA regulations do not, however, address genetic or any other health information created by researchers or disclosures by researchers. This is one reason why legislation is warranted to reduce the privacy risks described by Lin *et al.* But it would have to include much more than simply additional penalties for violating HIPAA rules. It would have to extend privacy protections to genetic information wherever it is created or maintained in an individually identifiable form. The HIPAA regulations cannot do this, and neither can state antidiscrimination laws that only regulate employers and health insurers (*3*). As my colleagues and I have argued elsewhere, it can only be achieved by a comprehensive federal genetic privacy law (*4*).

PATRICIA A. ROCHE
Department of Health Law, Bioethics and Human Rights, Boston University School of Public Health, 715 Albany Street, Talbot 3 West, Boston, MA 02118, USA. E-mail: pwroche@bu.edu

References and Notes
1. Regulations promulgated by the U.S. Department of Health and Human Services under the Health Insurance Portability and Accountability Act (HIPAA) of 1996, codified at 45 CFR § 164.500 et seq.
2. The rules also apply to health information held by employers and "health care clearinghouses." 45 CFR § 164.501.
3. Examples of state laws governing genetic privacy can be found at www.genome.gov/PolicyEthics/LegDatabase/PubSearchResult.cfm.
4. P. Roche, G. Annas, *Nature Rev. Genet.* **2**, 392 (2001).

THE NEWS FOCUS ARTICLE "PRIVACY RULE creates bottleneck for U.S. biomedical researchers" (J. Kaiser, 9 July 2004, p. 168) and the Policy Forum "Genomic research and human subject privacy" (Z. Lin *et al.*, 9 July 2004, p. 183) discuss the difficulties of exploiting new genetic capabilities for medical research while also protecting individuals' privacy. Lin *et al.* argue that, lacking technological solutions, this issue can only be treated through public policy regulations, and Kaiser highlights the problems that such regulations can cause for researchers.

An alternative solution exists. Well-established cryptographic techniques can

Letters to the Editor

Letters (~300 words) discuss material published in *Science* in the previous 6 months or issues of general interest. They can be submitted through the Web (www.submit2science.org) or by regular mail (1200 New York Ave., NW, Washington, DC 20005, USA). Letters are not acknowledged upon receipt, nor are authors generally consulted before publication. Whether published in full or in part, letters are subject to editing for clarity and space.

allow researchers to ask arbitrary sets of questions about personal data without them being able to discover anything beyond the aggregate answer they seek. This solution, which relies on zero-knowledge cryptographic techniques developed in the context of secure distributed computation, does not require a trusted third party or regulatory enforcement. A variation of this mechanism (*1*) allows a researcher to issue a survey to a number of individuals who can answer in an anonymous fashion, but in such a way that they can still be tracked over time and queried on additional items without the researcher learning their identity. These approaches offer an alternative solution to a widespread problem that others have so far argued can only be addressed through regulations and their enforcement.

BERNARDO A. HUBERMAN AND TAD HOGG
HP Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, USA.

Reference

1. B. A. Huberman, T. Hogg, *Nature Biotechnol.* **20**, 332 (2002).

Response

WE AGREE THAT A COMPREHENSIVE FEDERAL genetic privacy law could be very helpful in protecting patient privacy as well as the research mission. We need to point out, however, that many genetic research studies use samples obtained during the course of medical care—collected by health providers who certainly fall under the HIPAA umbrella. Thus, we do not believe that HIPAA is sufficiently clear in defining the boundaries between clinical uses and research uses of DNA sequence data, and therein lies our concern.

Protected health information (PHI) defined in HIPAA certainly includes genetic information. However, to use PHI for research, the rules do not explicitly declare genetic information as a restricted biometric identifier that must be removed from de-identified data sets. If this information were to be recognized as biometric identifiers, or were to be put under the category of "any unusual characteristics that may disclose the identity of a human subject," it could create major obstacles for research use. In particular, stripping identifiable DNA sequence information to create de-identified data sets could render genomic data virtually useless for research, because this information is precisely that required to perform genetic studies of how sequence variations correlate with inheritable diseases.

In the end, however, we agree with Roche that these issues should be clarified at the federal level.

Although we have not ruled out technical solutions to these problems, we do not think that cryptographic techniques, as suggested by Huberman and Hogg, by themselves are a sufficient solution. Cryptographic methods make it hard for people to intercept information, but they do not prevent authorized people from gathering sufficient data to re-identify individuals. For instance, if researchers received patients' social security numbers (SSNs) along with medical data, the subjects would be easily identifiable, even if cryptographic protection were used in conveying the data. Even a very small number of independent single nucleotide polymorphisms (SNPs) identifies an individual almost as precisely as an SSN does. Researchers with access to a large number of SNPs and corresponding phenotype data could therefore be in a position to identify some individuals.

> **"[S]tripping identifiable DNA sequence information ... could render genomic data virtually useless for research..."**
>
> –*ALTMAN* ET AL.

RUSS B. ALTMAN,[1] ZHEN LIN,[1] ART B. OWEN[2]
[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305–5120, USA. [2]Department of Statistics, Stanford University, Stanford, CA 94035–4065, USA.

Autism and Deficits in Attachment Behavior

A. MOLES *ET AL.* REPORTED A "DEFICIT IN attachment behavior in mice lacking the μ-opioid receptor gene" (Reports, 25 June 2004, p. 1983). They stated that their results "may indicate a molecular mechanism for diseases characterized by deficits in attachment behavior, such as autism," an assumption recapitulated in Mary Beckman's news story "The mice that don't miss mom: love and the μ-opioid receptor" (News of the Week, 25 June 2004, p. 1888) and other news coverage.

However, the basis for D'Amato *et al.*'s speculation that autism is "characterized by deficits in attachment behavior" was a 25-year-old theoretical treatise (*1*), which subsequently failed to be supported by numerous empirical investigations. In 1984, Sigman and colleagues first demonstrated that young children with autism behaved no differently in their response after separation from and reunion with their primary caregiver than did children with other developmental disabilities (*2*). Following Sigman's seminal study, every laboratory experiment investigating attachment behavior with young children with autism has replicated the initial finding, demonstrating unambiguously that children with autism are as securely attached to their mothers as are their peers (*3*–*8*).

Given the striking lack of empirical evidence to support D'Amato and colleagues' speculation that their results are relevant to autism, we urge considerably greater caution.

MORTON A. GERNSBACHER,[1] CHERYL DISSANAYAKE,[2] H. HILL GOLDSMITH,[1] PETER C. MUNDY,[3] SALLY J. ROGERS,[4] MARIAN SIGMAN[5]
[1]University of Wisconsin at Madison, Madison, WI 53706–1611, USA. [2]La Trobe University, Bundoora, Victoria 3083, Australia. [3]University of Miami, Coral Gables, FL 33124–0721, USA. [4]University of California Davis Medical Center, Sacramento, CA 95817, USA. [5]University of California, Los Angeles, Los Angeles, CA 90024–1759, USA.

References

1. J. Panksepp, *Trends Neurosci.* **2**, 174 (1979).
2. M. Sigman *et al.*, *J. Autism Dev. Disord.* **14**, 231 (1984).
3. T. Shapiro, M. Sherman, G. Calamari, D. Koch, *J. Am. Acad. Child Adolesc. Psychiatry* **26**, 480 (1987).
4. M. Sigman, P. Mundy, *J. Am. Acad. Child Adolesc. Psychiatry* **28**, 74 (1989).
5. S. J. Rogers, S. Ozonoff, C. Maslin-Cole, *J. Am. Acad. Child Adolesc. Psychiatry* **32**, 1274 (1993).
6. L. Capps *et al.*, *Dev. Psychopathol.* **6**, 249 (1994).
7. C. Dissanayake, S. A. Crossley, *J. Child Psychol. Psychiatry* **37**, 149 (1996).
8. C. Dissanayake, S. A. Crossley, *J Autism Dev. Disord.* **27**, 295 (1997).

Response

GERNSBACHER AND COLLEAGUES CLAIM THAT our findings are not relevant to autism because autism is not characterized by deficits in mother-child attachment. To support their argument, these authors cite several studies showing that children with autism can form a secure attachment with their mothers.

Mice lacking the μ-opioid receptor gene are not a model of autism or a model of any other human clinical syndrome as described in current psychiatric classification systems. Rather, our findings could be useful for identifying a biological pathophysiology common to a variety of conditions that are currently classified in very different categories of psychiatric nosography. Current classification of psychiatric disorders is not based on measures of the underlying genetic or biological pathophysiology of the disorders (*1*). As a consequence, diagnostic categories often include heterogeneous populations, and their classification is not of great help in deciding if a core neurobehavioral deficit is present or not in a psychopathological condition.

The question of whether autistic children are able to form an attachment bond (secure or insecure) with their mothers cannot be answered unequivocally, as indicated by a

recent review by Rutgers *et al.* (*2*). That a substantial proportion of autistic children (as diagnosed by current classification criteria) may develop secure attachments is not in contradiction with the clinical notion that many children and adults with autism show a marked reduction in the desire to engage others socially and that the opioid system could be implicated in the pathogenesis of these social deficits (*3*). Autism is not only a heterogeneous condition but is also best described as a continuum rather than as a category (especially its core component of defective reciprocal social behavior) (*4*).

In conclusion, we believe that our findings in mice lacking the μ-opioid receptor gene should be interpreted as an attempt to model a biological dysfunction that can be present in some (but not in all) cases of autism, as well as in other psychiatric conditions that have in common a deficit in the capacity to experience affiliative reward (*5*). Such conditions might include schizophrenia spectrum disorders and primary psychopathy (*6*).

FRANCESCA D'AMATO AND ANNA MOLES
Department of Psychobiology and Psychopharmacology, Consiglio Nazionale delle Ricerche Institute of Neuroscience, Viale Marx 43, 00137 Rome, Italy.

References

1. I. I. Gottesman, T. D. Gould, *Am. J. Psychiatry* **160**, 636 (2003).
2. A. H. Rutgers *et al.*, *J. Child Psychol. Psychiatry* **45**, 1123 (2004).
3. L. Sher, *Med. Hypotheses* **48**, 413 (1997).
4. J. R. Costantino, R. D. Todd, *Arch. Gen. Psychiatry* **60**, 524 (2003).
5. R. A. Depue, J. V. Morrone-Strupinsky, *Behav. Brain Sci.*, in press.
6. A. Troisi, F. R. D'Amato, *Behav. Brain Sci.*, in press.

TECHNICAL COMMENT ABSTRACTS

COMMENT ON "Epitaxial $BiFeO_3$ Multiferroic Thin Film Heterostructures"

W. Eerenstein, F. D. Morrison, J. Dho, M. G. Blamire, J. F. Scott, N. D. Mathur

Wang *et al.* (Reports, 14 March 2003, p. 1719) reported $BiFeO_3$ films as promising multiferroic materials, with impressive ambient ferroelectric polarizations and magnetizations due to epitaxial strain. However, fully oxygenated films with Fe^{3+} possess no significant magnetization. The large magnetization observed by Wang *et al.* is attributable to $Fe^{2+/3+}$, and the concomitant leakage precludes quantitative polarization measurements or practical devices.
Full text at www.sciencemag.org/cgi/content/full/307/5713/1203a

RESPONSE TO COMMENT ON "Epitaxial $BiFeO_3$ Multiferroic Thin Film Heterostructures"

J. Wang, A. Scholl, H. Zheng, S. B. Ogale, D. Viehland, D. G. Schlom, N. A. Spaldin, K. M. Rabe, M. Wuttig, L. Mohaddes, J. Neaton, U. Waghmare, T. Zhao, R. Ramesh

We report results of additional experiments that bear on the issues raised in the comment by Eerenstein *et al.* Superconducting quantum interference device magnetometry reveals a magnetic moment of ~70 electromagnetic units (emu)/cm^3, which decreases with film thickness to about 8 to 10 emu/cm^3, a finding confirmed by x-ray magnetic circular dichroism studies. The possible role of oxygen vacancies and the consequent formation of Fe^{2+} on the origin of the magnetism is discussed.
Full text at www.sciencemag.org/cgi/content/full/307/5713/1203b

CORRECTIONS AND CLARIFICATIONS

Editors' Choice: "Nothing in common (7 Jan., p. 18). This paragraph should have been attributed to ShJS.

Editors' Choice: "Of mice..." (24 Dec 2004, p. 2164). This paragraph should have been attributed to ShJS.

Comment on "Epitaxial $BiFeO_3$ Multiferroic Thin Film Heterostructures"

Wang *et al.* (*1*) recently reported multiferroic behavior, with ferromagnetic and ferroelectric polarizations that are both large at room temperature, in thin strained films of $BiFeO_3$ (BFO). Although at room temperature, bulk BFO is ferroelectric (*2*) and antiferromagnetic (*3*–*5*), Wang *et al.* (*1*) reported that a 70-nm film shows both an enhanced ferroelectric polarization (90 μC cm^{-2}) and a substantial magnetization (1 μ_B/Fe). This remains the only report of a robust room-temperature multiferroic and suggests the potential for novel devices that exploit the anticipated strain-mediated magnetoelectric coupling between the two ordered ground states. In this Comment, we argue that epitaxial strain does not enhance the magnetization and polarization in $BiFeO_3$.

Like Wang *et al.* (*1*), we grew BFO films on 50-nm underlayers of $SrRuO_3$ (SRO) on $SrTiO_3$ (001) substrates (STO). In addition, we used plain STO and conducting 0.2% atomic Nb-doped $SrTiO_3$ substrates (Nb-STO). Both BFO and SRO films were grown by pulsed laser deposition with a KrF excimer laser (248 nm, 1 Hz, target-substrate distance = 8 cm). BFO films were grown (670°C, 8 Pa O_2, 1.6 J cm^{-2}) using a Bi-rich target of $Bi_{1.2}FeO_3$, because Bi is volatile (*6*). SRO films were grown (650°C, 9 Pa O_2, 1.7 J cm^{-2}) using a stoichiometric target. The growth rate for both BFO and SRO was 10 Å/min. After deposition, films were cooled at 5°C/min to 400°C in 40 kPa oxygen, annealed for 1 hour, and then cooled to room temperature at 8°C/min.

The crystalline quality of our films was investigated with high-resolution x-ray diffraction (Fig. 1). The reciprocal space maps show that the BFO and SRO in-plane lattice parameters are equal to the STO lattice parameter of 3.905 Å, consistent with coherently strained films. It should be noted that Wang *et al.* (*1*) performed their calculations of saturation polarization using the bulk SRO lattice parameter of 3.935 Å for the BFO in-plane lattice parameter. In our work, BFO films as thin as 40 nm gave equivalent x-ray results, but 300-nm BFO films were found to be relaxed. All coherently strained films showed a root-mean-square surface roughness of 2 nm, as determined by atomic force microscopy.

BFO film stoichiometry was determined from quantitative energy dispersive x-ray spectroscopy in a thick (~400 nm) film grown at a laser repetition rate of 2 Hz. The Fe:Bi ratio was found to be unity, within the error of the technique of a few percent. The Fe oxidation state was investigated for BFO/STO samples grown at 1 Hz and 2 Hz with x-ray photoelectron spectroscopy. All samples behaved similarly; a representative scan of the Fe 2p line is shown in Fig. 2. The position of this line is expected to be 711 eV for Fe^{3+} and 709.5 eV for Fe^{2+}, and the position of the satellite is expected at 719 eV for Fe^{3+} and 716 eV for Fe^{2+} (*7*). From Fig. 2, we deduce that the oxidation state of Fe in our BFO/STO films is Fe^{3+} and that there is no evidence for Fe^{2+} within a resolution of a few atomic percent.

To verify that our BFO films are insulating and ferroelectric, we performed piezoresponse microscopy (*8*) on BFO/Nb-STO and

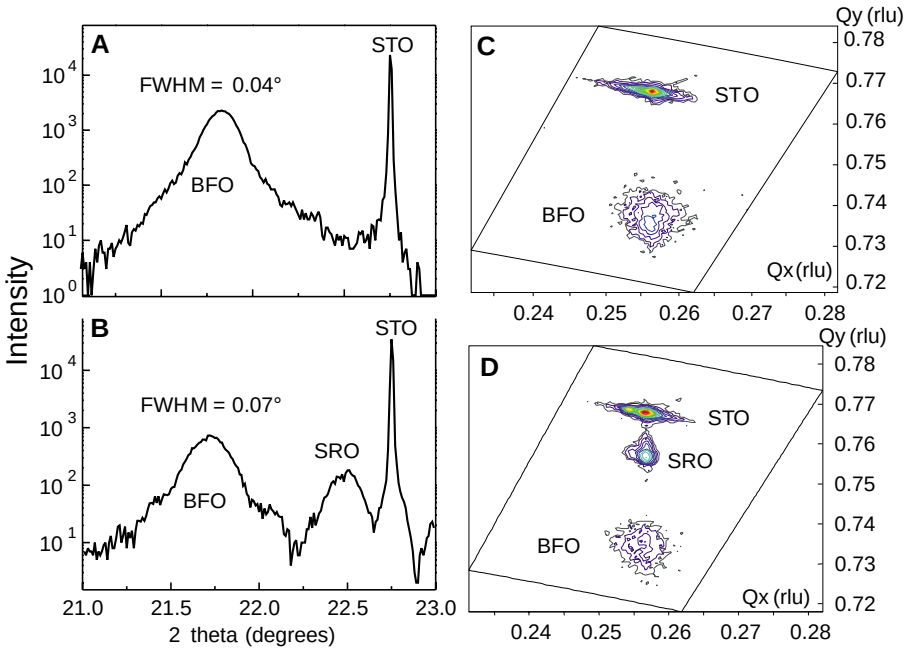
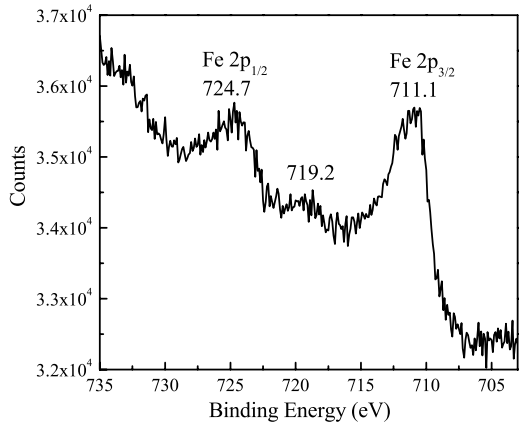


Fig. 1. Ω-2θ scan of the (001) pseudocubic reflection for (**A**) BFO(80 nm)/Nb-STO and (**B**) BFO(90 nm)/SRO(50 nm)/STO. The corresponding reciprocal space maps of the (013) pseudocubic reflections are shown in (**C**) and (**D**). Reciprocal lattice units (rlu) Q_x and Q_y are equal to the inverse lattice spacing. All in-plane film lattice parameters can be seen to be equal to the STO lattice parameter of 3.905 Å. The BFO out-of-plane lattice parameter in both samples is 4.07 Å, and the corresponding value for SRO is 3.95 Å.

Fig. 2. X-ray photoelectron spectrum of the Fe 2p lines for BFO(40 nm)/STO.



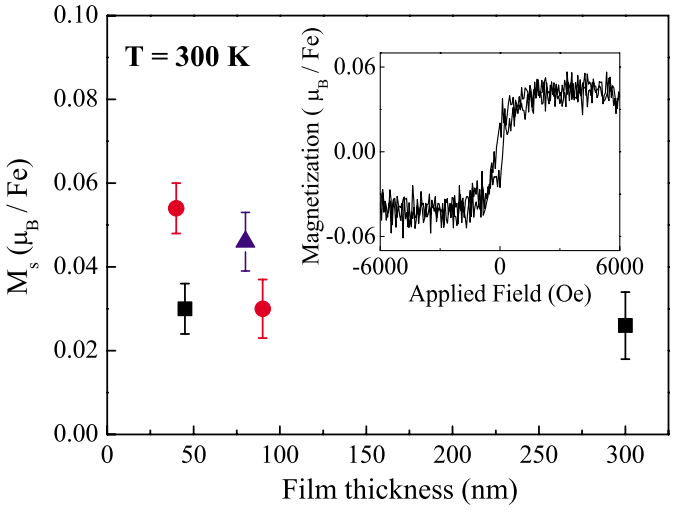


Fig. 3. Saturation magnetization M_s of BFO films versus film thickness. Black squares, BFO/STO; red circles, BFO/SRO(50 nm)/STO; blue triangle, BFO/Nb-STO. The inset shows a typical hysteresis loop. Magnetometer axis in-plane and parallel to STO [100]. Before taking magnetic measurements, we used emery paper to grind away silver dag from the sides and undersides of the substrates to eliminate spurious magnetic signals that we attribute to material originating from the heater block.

BFO/SRO/STO, and impedance spectroscopy using sputter-deposited Pt on BFO/Nb-STO. The former technique confirmed ferroelectric switching. The latter technique showed our films to be low-loss (2% loss tangent in 10 kHz to 1 MHz) and nonconducting (resistivity $>10^{10}$ Ωcm). The high-temperature conductivity data showed an activation energy of 1.03 ± 0.05 eV, compatible with only a small concentration of oxygen vacancies (*9*, *10*). In agreement with the literature (*11*), the effective dielectric constant (10 kHz to 1 MHz) was found to be 70 ± 2 at ambient temperature, increasing to ~375 at 550 K.

Magnetic measurements were taken at room temperature with a Princeton Measurements Corporation (Princeton, NJ) vibrating sample magnetometer. In Fig. 3, we plot saturation magnetization M_s as a function of BFO film thickness. BFO films both with and without the SRO underlayer behave similarly, which precludes any substantial magnetic contribution from SRO. All BFO films show an essentially thickness-independent M_s that is less than 0.06 μ_B/Fe, which rules out the strain-enhanced magnetization inferred by Wang *et al.* (*1*). Indeed, strain should not modify antiferromagnetic exchange interactions between singly occupied Fe^{3+} *3d* orbitals. We note that density functional calculations (*12*) of unstrained BFO suggest a local magnetization of around 0.05 μ_B/Fe, consistent with the findings presented here.

The small M_s that we observed is reminiscent of bulk behavior (*3*–*5*). The large value observed by Wang *et al.* (*1*) could arise as a result of a substantial Fe^{2+} fraction. If this fraction were 50%, then M_s = 0.5 μ_B/Fe, and the composition would be $BiFeO_{2.75}$. However, this mixed-valent system with oxygen vacancies would be expected to possess an electrical resistivity that is orders of magnitude lower than the figure of ~10^9 Ω cm stated in (*1*). In this scenario, the reported increase of polarization with decreasing film thickness (*1*) could be an experimental artifact. Regarding the polarization (50 to 60 μC cm^{-2}) of the thicker (200 nm) films measured by Wang *et al.*, although it is larger than the value recorded (*2*) in a poor-quality bulk sample (6 μC cm^{-2}), it is, in effect, not epitaxially enhanced with respect to unstrained polycrystalline films (40 μC cm^{-2}) (*13*). Therefore, we conclude that ferroelectric polarization is not enhanced by strain in BFO films. Indeed, calculations that supersede those presented by Wang *et al.* show 90 μC cm^{-2} in unstrained BFO (*14*).

We conclude that an increased thickness-dependent magnetization is not an intrinsic property of fully oxygenated and coherently strained epitaxial BFO films that exhibit a high electrical resistivity. If it is only possible to achieve substantial magnetization values in deoxygenated BFO, then the applications potential is reduced, because an increased electrical conductivity will be detrimental to ferroelectric performance.

W. Eerenstein*
Department of Materials Science
University of Cambridge
Pembroke Street, Cambridge, CB2 3QZ, UK

F. D. Morrison
Department of Earth Sciences
University of Cambridge
Downing Street, Cambridge, CB2 3EQ, UK

J. Dho* and *M. G. Blamire
Department of Materials Science

J. F. Scott
Department of Earth Sciences

N. D. Mathur
Department of Materials Science
**To whom correspondence should be addressed.*
E-mail: we212@cam.ac.uk

References and Notes

1. J. Wang *et al.*, *Science* **299**, 1719 (2003).
2. J. R. Teague, R. Gerson, W. J. James, *Solid State Commun.* **8**, 1073 (1970).
3. S. V. Kiselev, R. P. Ozerov, G. S. Zhanov, *Sov. Phys. Dokl.* **7**, 742 (1963).
4. G. A. Smolenskii, V. M. Yudin, E. S. Sher, Y. E. Stolypin, *Sov. Phys. JETP* **16**, 622 (1963).
5. I. Sosnovska, T. Peterlin-Neumaier, E. Steichele, *J. Phys. C Solid State Phys.* **15**, 4835 (1982).
6. A. Garg, S. Dunn, Z. H. Barber, *Integrat. Ferroelectr.* **31**, 13 (2000).
7. Th. Schedel-Niedrig, W. Weiss, R. Schlögl, *Phys. Rev. B* **52**, 17449 (1995).
8. A. Gruverman, O. Auciello, H. Tokumoto, *Annu. Rev. Mater. Sci.* **28**, 101 (1998).
9. J. F. Scott, *Ferroelectric Memories* (Springer, Berlin, 2000).
10. S. Zafar, R. E. Jones, B. Jiang, B. White, P. Chu, D. Taylor, S. Gillespie, *Appl. Phys. Lett.* **73**, 175 (1998).
11. M. Mahesh Kumar, V. R. Palkar, K. Srinivas, S. V. Suryanarayana, *Appl. Phys. Lett.* **76**, 2764 (2000).
12. C. Ederer, N. A. Spaldin, *Phys. Rev. B* **71**, 060401 (2005).
13. K. Y. Yun, M. Noda, M. Okuyama, *Appl. Phys. Lett.* **83**, 3981 (2003).
14. J. B. Neaton, C. Ederer, U. V. Waghmare, N. A. Spaldin, K. M. Rabe, *Phys. Rev. B* **71**, 014113 (2005).
15. We are grateful to A. Garg and Z. H. Barber for advice on bismuth-excess ablation targets; M. E. Vickers for help with the x-ray data; T. Hibma for x-ray photoelectron spectroscopy facilities in the Department of Physical Chemistry, University of Groningen, The Netherlands; P. S. Roberts, M. Alexe, and C. Harnagea for help with piezoimaging; and N. A. Spaldin, C. Ederer, and R. Ramesh for helpful discussions. This work was funded by the Royal Society, the UK Engineering and Physical Sciences Research Council, an EU Marie Curie Fellowship (W.E.), and a Korea Science and Engineering Foundation postdoctoral fellowship (J.D.).

7 July 2004; accepted 29 December 2004
10.1126/science.1105422

Response to Comment on "Epitaxial $BiFeO_3$ Multiferroic Thin Film Heterostructures"

In this response, we report the results of additional experiments that bear on the issues raised in the comment by Eerenstein *et al.* (*1*) and suggest some additional possible reasons for the results they have obtained.

In Wang *et al.* (*2*), we reported on the ferroelectric and magnetic properties of $BiFeO_3$ (BFO) epitaxial thin films. Our results demonstrated a thickness dependence in these properties, and we suggested that a likely explanation of these effects was that heteroepitaxial strain induced a monoclinic distortion, relaxing gradually with increasing thickness. Detailed x-ray studies (*3*) have shown evidence for such a monoclinic structure, as well as no evidence for secondary phases. The out-of-plane lattice parameter for the BFO layer progressively increases as the thickness is decreased, consistent with the expected effect of epitaxial constraint.

In contrast to all prior studies, our epitaxial BFO thin films showed a large spontaneous polarization. We considered the hypothesis that this change could be directly attributed to the epitaxially induced change in structure. Our theoretical discussion also included the possibility that the large polarization in thin films could be the result of a change in the switching mechanism in thin films compared with the bulk (*4*), a change that could be related to other differences as well as to epitaxial strain. Another possibility, originally suggested by Teague *et al.* (*5*), is that high leakage in the bulk samples, somehow reduced in the films, could have prevented prior researchers (over the past four decades) from observing the large value of spontaneous polarization of BFO. These latter proposals are consistent with various subsequent measurements of high polarization of BFO in thin film form (*6*–*8*).

Our report (*2*) also included larger values for the magnetization of very thin films (thickness less than ~100 nm), again different from the previously reported magnetic properties of the bulk material. We have continued to investigate the magnetic behavior of the films; Fig. 1A summarizes our recent magnetic measurements from a series of films with systematically varying thickness. In contrast to our original measurements, which were carried out using a vibrating sample magnetometer (VSM), we have now been able to measure our films using a more sensitive, higher resolution superconducting quantum interference device (SQUID) magnetometer (Fig. 1B), which was not accessible to us at the time of preparation of the original paper. Measured using this instrument, the very thin films show a magnetic moment of about 70 to 80 electromagnetic units (emu)/cm^3 (corresponding to a magnetization of about 0.5 μ_B/formula unit), which progressively decreases as the film thickness is increased to above 120 nm. Although smaller than in the original report, this value is still much greater than those reported by Eerenstein *et al.* (*1*).

In these samples, detailed transmission electron microscopy studies of both planar and cross sections did not reveal any second phases. Piezoimaging with an atomic force microscope (AFM) has clearly shown the existence of piezoelectricity down to at least 30-nm thickness, although the degree of leakage goes up with decreasing thickness. X-ray diffraction studies have revealed no secondary phases; we have used Rutherford backscattering and energy dispersive x-ray spectroscopy to analyze the cation composition and see a Bi:Fe ratio of 1:1 in the films, within experimental error. However, the oxygen stoichiometry is difficult to determine exactly. As is common in the oxide thin film literature, we have used nominal oxygen composition to describe the film—in this case, $BiFeO_3$. Such notation does not necessarily, and in this case was not intended to, imply that the films are free of point defects, including vacancies, interstitials, or antisite defects.

We have been pursuing approaches to understand the origin of higher moments in the very thin films using x-ray absorption spectroscopy (XAS) and x-ray magnetic circular dichroism (XMCD) measurements with a magnetic field of 800 Oe [details of the measurements are given in (*9*)]. Fig. 2A shows XMCD results from the same series of films that were used to obtain the structural and magnetic data above. The signal magnitude, normalized to the total XAS signal, is a measure of the magnetic moment in the sample. The data clearly show that the 30-nm-thick film has a distinct magnetic signal, which progressively decreases as the film thickness is increased. The position of the three peaks also reveals the mixed oxidation state of the sample by comparison with a known sample (Fe_3O_4).

Fig. 2B compares the normalized XMCD spectrum of the 30-nm BFO film to that of a Fe_3O_4 film. Several key points emerge from this comparison. First, using the Fe_3O_4 spectrum as a fingerprint, we can clearly identify the peak at ~709.5 eV as the Fe^{+2} peak, consistent with our x-ray photoelectron spectroscopy results. Second, this peak progressively vanishes as the thickness is increased (Fig. 2A). Third, the total intensity of the spectrum from the 30-nm BFO film is approximately 15% of that from the Fe_3O_4 film, which suggests that the total moment is also approximately 15% that of Fe_3O_4. Finally, the ratio of the three peaks in Fig. 2B is not the same as that in Fe_3O_4, for which detailed studies have revealed a ratio of 1:1:1 for the Fe^{2+}(oct):Fe^{3+}(tet):Fe^{3+}(oct). What is interesting, and puzzling, is that the +2 state vanishes at thicknesses higher than

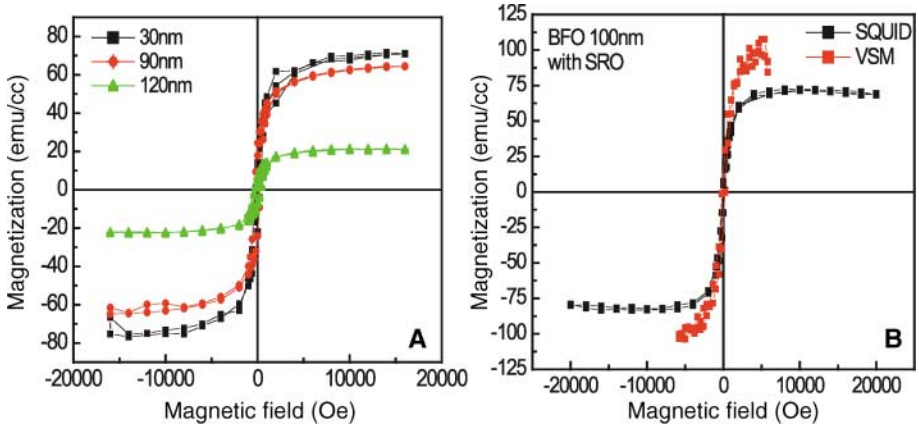


Fig. 1. (**A**) Magnetic response from BFO films grown on 001 STO with different thickness. (**B**) Comparison of VSM and SQUID results on the 100-nm BFO/SRO/STO (001) sample used in (*2*). Care was taken to remove any silver spots on the substrate by grinding with emery paper.

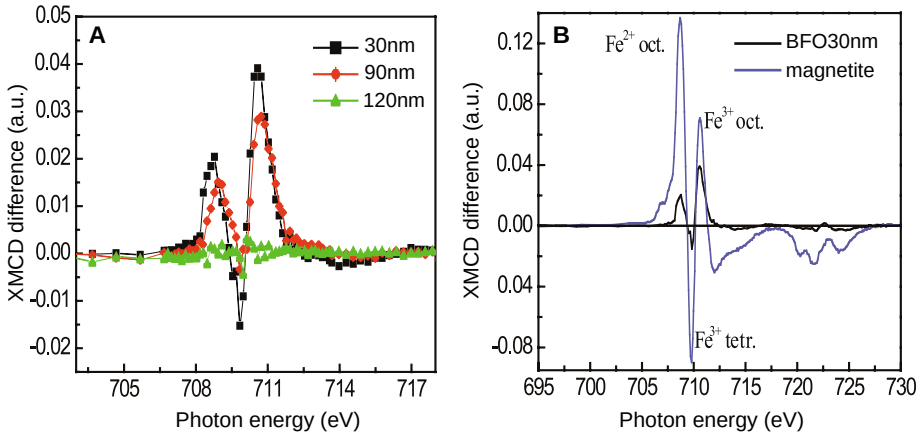


Fig. 2. (**A**) XMCD results for the 30-nm, 90-nm, and 120-nm BFO/STO (001) samples. (**B**) Comparison of XMCD spectrums of Fe_3O_4 and the BFO film.

~100 nm, concomitant with the magnetism decreasing to the level of 8 to 10 emu/cm^3.

The most likely origin of Fe^{2+} in the thin films is the presence of oxygen vacancies, which are typically quite common in perovskites. Our films were grown under relatively reducing conditions, compared with those of Eerenstein *et al*. (*1*), and thus favored the formation of Fe^{2+}. In the absence of Fe^{2+}, our observations (that is, at thicknesses >120 nm) and theoretical analysis (*10*) suggest that the magnetization should be ~8 to 10 emu/cm^3, consistent with the observations of Eerenstein *et al*. (*1*).

With the presence of Fe^{2+} in the films, a large number of possible models could be invoked to explain the extra observed moment. We summarize the most plausible here but emphasize that we are still actively working both experimentally and theoretically to elucidate the origin of the magnetic behavior we have observed.

One possibility is a ferrimagnetic arrangement in which the moments of the Fe^{2+} ions are aligned oppositely to those of the Fe^{3+} ions, leading to a net magnetic moment. Eerenstein *et al*. (*1*) suggest that an oxygen count of 2.75 (instead of 3) would be required to produce our experimentally observed moment of 0.5 μB/Fe. This would be the case for high-spin Fe^{2+}, but the oxygen deficiency required to produce the observed moment would be much lower if low-spin Fe^{2+} were present. We suggest that such a ferrimagnetic arrangement may be unlikely; even if Fe^{2+} and Fe^{3+} were to couple ferrimagnetically locally, they would need to also have long-range ordering or the moments from locally ferrimagnetic regions would cancel out. An alternative possible mechanism is a gradual increase in the canting angle as the thickness is reduced, which could be driven by increasing distortions resulting from epitaxial strain, oxygen vacancies, or both. Clearly, this aspect is scientifically very interesting and requires further experimental and theoretical study.

J. Wang
Department of Materials Science and Engineering
University of Maryland
College Park, MD 20742, USA

A. Scholl
Advanced Light Source
Lawrence Berkeley Laboratory
Berkeley CA, 94720, USA

H. Zheng
Department of Materials Science and Engineering
University of Maryland, College Park

S. B. Ogale
Department of Physics
University of Maryland, College Park

D. Viehland
Department of Materials Science and Engineering
Virginia Polytechnic Institute
Blacksburg, VA 24061, USA

D. G. Schlom
Department of Materials Science and Engineering
Pennsylvania State University
University Park, PA 16802–5055, USA

N. A. Spaldin
Materials Department
University of California
Santa Barbara, CA 93106–5050, USA

K. M. Rabe
Department of Physics and Astronomy
Rutgers University
Piscataway, NJ 08854, USA

M. Wuttig
Department of Materials Science and Engineering
University of Maryland, College Park

L. Mohaddes
Department of Materials Science and Engineering
and Department of Physics
University of California
Berkeley, CA 94720

J. Neaton
Department of Materials Science and Engineering
and Department of Physics
University of California, Berkeley

U. Waghmare
Jawaharlal Nehru Center, Bangalore, India

T. Zhao
Department of Materials Science and Engineering
and Department of Physics
University of California, Berkeley

R. Ramesh
Department of Materials Science and Engineering
and Department of Physics
University of California, Berkeley
E-mail: rramesh@berkeley.edu

References

1. W. Eerenstein *et al.*, *Science* **307**, 1203 (2005); www.sciencemag.org/cgi/content/full/307/5713/1203a.
2. J. Wang *et al.*, *Science* **299**, 1719 (2003).
3. J. L. Wang, thesis, University of Maryland, College Park (2005).
4. J. B. Neaton *et al.*, *Phys. Rev. B* **71**, 014113 (2005).
5. J. R. Teague, R. Gerson, W. J. James, *Solid State Commun.* **8**, 1073 (1970).
6. K. Y. Yun *et al.*, *Appl. Phys. Lett.* **83**, 3981 (2003).
7. K. Y. Yun *et al.*, *J. Appl. Phys.* **96**, 3399 (2004).
8. K. Y. Yun *et al.*, *Jap. J. Appl. Phys.* **43**, L647 (2004).
9. P. Morrall *et al.*, *Phys. Rev. B* **67**, 214408 (2003).
10. C. Ederer, N. A. Spaldin, http://arxiv.org/abs/cond-mat/0407003.

11 August 2004; accepted 7 February 2005
10.1126/science.1103959

NEUROSCIENCE

A Piece of a Neuroscientist's Mind

Charles A. Nelson and Irving I. Gottesman

The Great Brain Debate
Nature or Nurture?
by John E. Dowling

Joseph Henry Press, Washington, DC, 2004. 197 pp. $24.95, C$32.95. ISBN 0-309-09223-X. Science Essentials. (*1*)

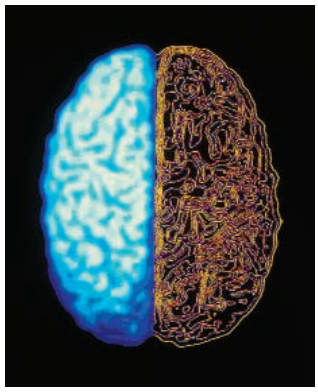
All sciences have their great debates, their burning controversies, their big unanswered questions. The increasingly coupled disciplines of neuroscience and psychology are no exception. Indeed, a surprising number of recent popular books touch on some of these questions, books such as Steven Pinker's *The Blank Slate* (*2*) and Gary Marcus's *The Birth of the Mind* (*3*). Both of these ambitious works tackle the archaic and divisive nature-nurture question, for example, although neither author is an expert in genetics, molecular biology, or neuroscience. That is not the case with *The Great Brain Debate*, as John Dowling is a first-rate visual neuroscientist who knows the brain and who inspiringly draws on his own research program to communicate with Carl Sagan–like clarity the complexities of the inner cosmos—brain function and brain development.

Not so long ago, the media was abuzz with early brain development—as in 1997, when the Clintons hosted a White House conference on the topic. The theme of the day was critical periods, especially the idea that the first three years of life represent a fixed window of opportunity for brain development. There the assumption was that good experiences led to "good brains" and bad experiences led to "bad brains." Leaving aside the fact that there was little scientific evidence to support this simplistic view, the conference and the media frenzy that followed did a masterful job of moving the study of brain development out of the laboratory into the real world. A case in point was Zell Miller's decision as governor of Georgia to send all newborns home from the hospital with a Mozart CD. Miller assumed that hearing Mozart would enhance early brain development, an assumption unfortunately without scientific support.

C. A. Nelson is at the Institute of Child Development, University of Minnesota, 51 East River Road, Minneapolis, MN 55455, USA. E-mail: canelson@umn.edu. I. I. Gottesman is in the Departments of Psychiatry and Psychology, University of Minnesota, Minneapolis, MN 55455, USA. E-mail: gotte003@umn.edu

Since 1997, there have been a plethora of books on this topic, and the MacArthur Foundation has underwritten an entire research network on early experience and brain development (www.macbrain.org). Many of the books were written for a lay audience; some, such as the National Academy of Sciences' *From Neurons to Neighborhoods* (*4*), were directed toward policy-makers and educators. Few were written by neuroscientists, and none attempted to deal with the complexity of gene-environment relations.

The nature-nurture debate dates back, at least, to the 17th-century philosophers Jean Jacques Rousseau and John Locke, who respectively argued that behavioral development was determined by unseen factors (of course, the gene hadn't been discovered yet) or by the environment. The debate continued through the 20th century, with some (such as John Watson in the 1920s) arguing that experience determined the course and form of virtually all aspects of behavioral development. Others (such as behavioral geneticist Thomas Bouchard and colleagues) used data from twins separated at birth to argue that genes play a very prominent role in everything from IQ to political beliefs to whether one is predisposed to divorce. We now realize that neither environmental not genetic determinism should have a role in contemporary efforts to understand and to explain ourselves.



The backdrop against which the debate has played out gradually changed as advances in neuroscience and in molecular biology began to infiltrate the field of psychology. Indeed, over the past few years, a number of landmark papers have been published that illustrate the complex relation between genes and environment. The pace of discovery has increased with recognition of the plasticity of the nervous system (including its ability to remodel itself in response to epigenetic programming driven by maternal behaviors), exciting reports of gene-environment interplay involving specific genes and specific environments in important human behaviors, and the ability to search bioinformatic databases online to find gene homologies between, say, *Homo sapiens* and zebrafish.

The book's title is a bit misleading, as Dowling devotes relatively little space to discussing this complex relation. And when he turns to it, he shifts from speaking as an insider to relying on secondary sources. Still, Dowling does a masterful job of laying out how the brain is born and then matures. His eloquent essay provides solid examples of what elements of brain development and brain function are under genetic control and which are largely guided by experience.

Among the sections on development and maturation, Dowling embeds discussions of some current controversies: notably the concept of critical periods, whether neurons continue to be made anew long after birth, and, conceptually linked to both, adult plasticity. He begins a consideration of how genes and experience can influence the course of brain and behavioral development. Alas, this topic (the subtitle of the book) is simply laid out without a great deal of elaboration, which is unfortunate because the author has clearly given it a great deal of thought. After discussing the aging brain (focused on neuropathology rather than normal aging), Dowling returns to the recast nature-nurture question in his final remarks. Despite the brevity of his exploration, there is just enough to whet the appetite and to stir the reader's imagination.

This highly readable account of brain development and brain function will appeal to an educated reader, one who is capable of critically evaluating the literature even if only from the perspective of an outsider. Our biggest complaint about *The Great Brain Debate* is that Dowling fails to tell us what he thinks about the roles of genes and experiences in brain development and maturation. Many scientists share his reticence, but he has earned the right to express his opinion rather than to simply lay out the data. In this context, one can only hope for a sequel.

References and Notes

1. The book may be read online, gratis, at http://books.nap.edu/catalog/11004.html.
2. S. Pinker, *The Blank Slate: The Modern Denial of Human Nature* (Viking, New York, 2002). Reviewed by P. Bateson, *Science* **297**, 2212 (2002).
3. G. Marcus, *The Birth of the Mind: How a Tiny Number of Genes Creates the Complexities of Human Thought* (Basic, New York, 2004). Reviewed by H. C. Barrett, *Science* **304**, 1601 (2004).
4. J. P. Shonkoff, D. A. Phillips, Eds., *From Neurons to Neighborhoods: The Science of Early Childhood Development* (National Academies Press, Washington, DC, 2000).

10.1126/science.1108869

RISK AND PUBLIC POLICY

Courting Disaster

Kenneth R. Foster

In this fine book, U.S. federal judge and public intellectual Richard A. Posner worries about events of very low probability but very high impact, such as possible human extinction. These are catastrophes of a different order than the usual hurricanes and floods that are subject to a growing academic literature on risk management (*1*).

Case in point: a high-velocity collision of a 2-kilometer asteroid with Earth might kill a billion people, whereas a 10-kilometer asteroid might extinguish our species. Even a 75-meter rock could release energy equivalent to a 100-megaton nuclear weapon, enough to obliterate a city. Thousands of such near-Earth objects are thought to exist. Some estimates place the yearly death toll averaged over megayears at 1500—more than the current annual number of fatalities from airline crashes.

Other potential catastrophes that Posner discusses include global climate change that leads to a sudden change in the Gulf Stream, (devastating agriculture in Western Europe) and widespread death from bioterrorism. Some of the threats border on science fiction, for example, the "gray goo" of a nanotechnology gone wild and the "strangelet scenario" of a high-energy physics experiment gone awry. But the interest in this book lies not in its doomsday scenarios, but in Posner's finely reasoned discussion, which reflects more legal and economic sophistication than is commonly found in the highly polarized debates on these issues. The 47 pages of references to legal, economic, and scientific papers show that he has done his homework on the subject.

Catastrophe
Risk and Response
by Richard A. Posner
Oxford University Press, New York, 2004. 332 pp. $28, £16.99. ISBN 0-19-517813-0.

For example, in a fine lawyerly analysis, Posner finds that experts who argue that human activity is causing global warming are more credible than naysayers who argue that it is not. He takes apart arguments of "doomsters" such as Paul Ehrlich, with their overly pessimistic predictions of environmental catastrophe, and technological optimists who minimize potential hazards of new technologies, such as John Maddox, former editor of *Nature*, and science journalist Gregg Easterbrook. He worries about legal justifications for "extreme measures," including curtailment of civil liberties and torture, in the fight against terrorism. He is less disturbed by the limited restrictions adopted by the United States since September 2001 than by the likely political consequences of another major terrorist attack. Posner might not be right, but his arguments have weight.

Relying on the dismal science for guidance, Posner monetarizes the anticipated costs of catastrophes and possible remediation strategies. Increased efforts to locate space rocks on collision course with Earth would be cost-effective, he finds.

Cost-benefit analysis is less supportive of caps on greenhouse gas emissions, as required by the Kyoto Protocol. The costs, particularly to the United States and other developed countries, would be huge in the short term while the benefits would accrue only in the very long run—unless, of course, a trigger point were reached that suddenly redirected the Gulf Stream. Expressed in terms of present value, Posner finds the costs to far outweigh the benefits. (By the same reasoning, the present value of a $1 million lottery award payable as a 20-year annuity is much less than $1 million). Posner favors, instead, a tax on carbon emissions to force the development of new technologies for clean fuels and economic methods to sequester carbon from the environment.

But cost-benefit analysis can only take one so far, given the extreme uncertainties surrounding the risks he considers. And for some risks, biological or nuclear terrorism for example, the likelihood of tragedy seems so high that cost-benefit analysis would not be needed to justify taking strenuous practical measures to counter them.

Posner will infuriate many scientists whom, he writes, have an "attitude gap created by the different goals, and resulting different mindsets, of science on the one hand and public policy on the other. The scientist qua scientist wants to increase scientific knowledge, not make the world safer—especially from science."

The strangelet scenario is a case in point. Shortly before a new high-energy accelerator was to begin operation at Brookhaven National Laboratory, a physicist raised concerns that a high-energy collision might trigger a runaway reaction that would quickly transform Earth into a 100-meter lump of inert matter. The lab director took the ethically dubious step of appointing an evaluation panel of physicists, all of whom had professional interests in seeing the experiments go forward. Posner dismisses as non sequiturs the various public statements by physicists intended to reassure the public of the improbability of the strangelet scenario. Seeing few economic benefits and a likely small but in fact unknown probability of disaster, he argues that high-energy research should be supported by universities rather than the government and that it should be brought under a strict regulatory umbrella.

This view, voiced by a judge with obviously high intelligence and sophistication, serves as a clear warning that scientists must pay attention to the social as well as to the technical dimensions of technological risk—and develop a better understanding of how nonscientists will interpret their pronouncements on the subject. Posner's perspective, very different from those held by most scientists, is a welcome addition to considerations of catastrophic risks.

Natural event with an impact. There is little doubt that a collision with an asteroid 10 kilometers in diameter, such as the one that hit the Yucatan 65 million years ago, would lead to the death of much of the world's population.

CREDIT: DON DAVIS/NASA

The reviewer is in the Department of Bioengineering, University of Pennsylvania, 220 South 33rd Street, Philadelphia, PA 19104–6392, USA. E-mail: kfoster@seas.upenn.edu

Reference

1. P. Grossi, H. Kunreuther, Eds., *Catastrophe Modeling: A New Approach to Managing Risk* (Springer, New York, 2005).

10.1126/science.1109699

SPACE SCIENCE

Crossroad for European Space Activity

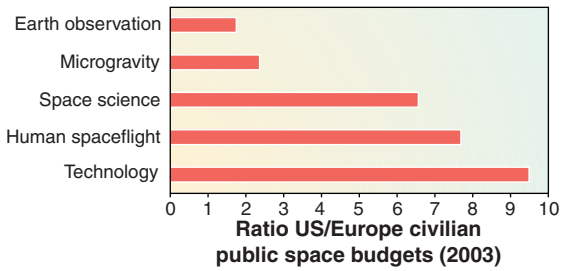
Bo Andersen

What is the future role of Europe in exploring space? European scientists are leaders in fields like astrometry with the Hipparcos mission (*1*) and the Gaia project to map our galaxy (*2*). The Ariane rockets launched more than half the telecommunication satellites over a period of more than a decade. The magnificent landing of the Huygens probe on Titan also demonstrates the technical and scientific capability of Europe.

However, despite successes, there are serious concerns. Europe has been lagging far behind the United States in spending; the total government space budgets for 2003 in the United States and Europe were US$34 and $6.2 billion, respectively (*3*). Russia has not had the economy or need to maintain the level of effort of the USSR.

Until recently, European public-funded space activities have been limited to national/bilateral projects and projects of the European Space Agency (ESA) (*4*). Public funding for space is predominantly through individual government contributions to the ESA (58%). Germany, France, Great Britain, and Italy provide most of the funding for the ESA, and their interests have generally determined ESA projects and objectives. In 2003, the European Union (EU) developed priorities for European space activities and stated that the current public funding levels were too low to maintain a competitive space industry and an independent European space launcher capability (*5*). Even including funding for the satellite navigation system Galileo, which the EU is developing with the ESA (*3*), the EU provides less than 5% of the public funding for European space investments. The EU would need to increase its investment in space exploration by a factor of 10 to really matter, and then only if it is done without impacting the member states' funding to ESA.

Although EU and ESA member states largely overlap, funding mechanisms and decision processes have been different and mostly uncoupled. This may change, as ESA and the EU have entered into a collaboration that will lead to closer coordination of budgets and division of responsibility (*6*).



Comparing the civilian space budgets in Europe and the United States (*2*).

The EU "White Paper" (*7*) gives upper and lower boundaries for budget increases. It was stated that the lower boundary is inadequate to maintain European capabilities. Even the more optimistic projected budget would have little effect on the gap between Europe and the United States. Any increases in EU budgets will have to come in connection with the ongoing 2007–13 budget discussions and should be decided by early 2006. However, several governments have capped their space budgets for the next few years. For example, France stated that their contributions to ESA will be below M 685 until 2009 (*8*). With the overall budgetary problems the EU has and will continue to have, even the lowest level of proposed growth does not appear realistic.

ESA or the EU wishes to expand in areas such as robotic space exploration of the solar system and human space exploration (*7*). Realistic funding of these initiatives can start in 2007 at the earliest; however, this requires a clear change of policy from all the European governments. We cannot meet costs of ongoing programs and expand without extra external funds or cannibalization of other fields like the space and earth sciences.

Exploration and the possible future colonization of the other habitable parts of our solar system will eventually be done through global cooperation and the importance of "joint space activities" has been emphasized in discussion of the new EU constitution (*9*). Europe has a long tradition in international collaboration with many different partners and could thus be an effective mediator and catalyst in these collaborations.

The EU White Paper has proposed development of human space flight capability (*7*). This could most efficiently be done in collaboration with Russia, which has a human exploration program and considerable experience. The plan to use Russian-built Soyuz rockets at the launch complex in French Guyana is a step toward closer collaboration. Further launcher development could be done in collaborations that maintain the strategic position of European industry.

The most promising new collaborations are in global monitoring. The European GMES (Global Monitoring for Environmental Security) initiative is being closely coordinated with the United States–initiated GEOSS (Global Earth Observation System of Systems). Recent meetings have been held to bring in developing nations as well (*10*). By coordinating planning and implementation, the space agencies can reduce costs, maximize results, and achieve global coverage more quickly than one or a few nations could.

Partnerships must be entered into from a position of strength; otherwise the European communities will be left with uninteresting tasks that have little strategic, scientific, or industrial importance. However, Europe should not embark on new endeavors unless there is a clear commitment for sufficient funding.

Europe has built a world-class program in spite of being on an economic diet. This diet has fostered creativity and efficiency. However, restriction can lead to starvation. The European space sciences communities need a healthy meal now, or else neither they nor the industries they support will survive.

The author is director of Space and Earth Sciences, Norwegian Space Centre, Oslo, Norway. E-mail: bo@spacecentre.no

References and Notes

1. http://astro.estec.esa.nl/Hipparcos/.
2. www.rssd.esa.int/index.php?project=Gaia.
3. "The European Space Sector in a Global Context: ESA's annual analysis 2003" (ESA BR-222, European Space Agency, Paris, 2004).
4. www.esa.int.
5. http://europa.eu.int/comm/space/off_docs_en.html.
6. www.esa.int/export/esaCP/Pr_62_2004_p_EN.html.
7. http://europa.eu.int/comm/space/whitepaper/index_en.html.
8. www.lemonde.fr/web/recherche_breve/1,13-0,37-800997,0.html.
9. http://europa.eu.int/comm/space/news/article_1280_en.html.
10. http://europa.eu.int/comm/space/russia/highlights/partnership_conference_en.html.

10.1126/science.1107684

NEUROSCIENCE

Making Synapses: A Balancing Act

Natasha K. Hussain and Morgan Sheng

In the brain, the massive network of neurons is connected at specialized junctions called synapses. Every aspect of nervous system function—coordinating movement, feeling pain, making decisions—involves neuronal communication across synapses in specific circuits of the brain. During development of the human brain, trillions of synapses, both excitatory and inhibitory, are constructed with extraordinary specificity as neurons grow and contact each other. How is this achieved? Several new studies published on page 1324 of this issue (*1*) and elsewhere (*2*, *3*) identify a key partnership of two cell surface molecules, β-neurexin and neuroligin, in the formation of both excitatory and inhibitory synapses.

Synapses are tiny junctions where an axon of a presynaptic neuron contacts a dendrite of a postsynaptic neuron (see the figure). The axon terminal releases a chemical signal (neurotransmitter) into the narrow synaptic cleft. Clustered on the postsynaptic membrane are receptor proteins that bind to and "receive" the neurotransmitter signal. Broadly speaking, synapses can be excitatory, with glutamate as the neurotransmitter, or inhibitory, with γ-aminobutyric acid (GABA) as the neurotransmitter. Both types of synapse exist side-by-side on the same postsynaptic neuron. For synapses to work, special proteins of the presynaptic and postsynaptic membranes must be not only aligned with each other, but also chemically matched (see the figure). A long-standing goal of neuroscience research has been to identify the proteins that mediate the physical apposition and chemical specificity of the pre- and postsynaptic membranes.

An attractive idea is that cell surface proteins binding to each other across the synaptic cleft (that is, "transsynaptically") could tether pre- and postsynaptic membranes. Moreover, if different pairs of adhesion molecules were found in chemically distinct synapses, their "lock-and-key" interactions could match the appropriate postsynaptic receptors to the presynaptic neurotransmitter. Hence the interest in β-neurexins and neuroligins: presynaptic and postsynaptic proteins, respectively, whose extracellular domains bind to each other (*4*).

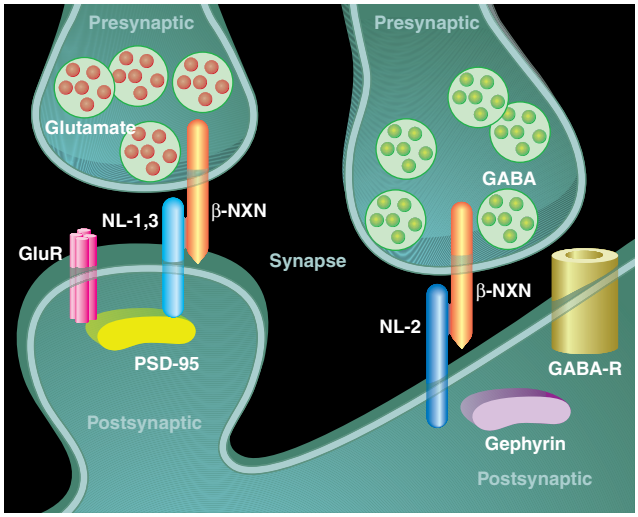
In addition to their ability to bridge pre- and postsynaptic membranes via their transsynaptic interactions, β-neurexins and neuroligins contain determinants in their intracellular regions that recruit the specific scaffolding proteins important for the assembly of pre- and postsynaptic protein complexes. Neuroligins are encoded by four related genes in humans. Neuroligin-1 (NL-1), when overexpressed on the surface of non-neural cells, can induce presynaptic differentiation in axons that come into contact with these cells. This effect depends on neuroligin's interaction with β-neurexin (*5*, *6*).

In their new study, Chih *et al.* (*1*) now demonstrate that overexpression of NL-1, -2, or -3 in cultured hippocampal neurons promotes the formation of excitatory postsynaptic specializations containing receptors for the excitatory neurotransmitter glutamate and for PSD-95, an intracellular scaffold protein that binds to glutamate receptors and neuroligin. These postsynaptic sites were aligned with presynaptic axon terminals that release glutamate. The assembly of such synaptic junctions required the extracellular interaction of neuroligin with β-neurexin as well as neuroligin's intracellular interaction with PSD-95 (see the figure).

Unexpectedly, the overexpression of NL-1, -2, or -3 in hippocampal neurons also stimulated the formation of more inhibitory synapses (*1*). The domain requirements for this effect of neuroligin were not examined. Thus, it is unclear whether the inhibitory synapses were formed by a mechanism similar to that inducing formation of excitatory synapses. When the protein levels of neuroligins in neurons were suppressed by RNA interference, the formation of both excitatory and inhibitory synapses was blocked (*1*). These "loss-of-function" experiments establish that endogenous neuroligins are critical for synapse formation during development.

In a complementary study published in a recent issue of *Cell*, Graf *et al.* showed that β-neurexin expressed on the surface of non-neural cells or attached to synthetic beads could induce the formation of excitatory and inhibitory postsynaptic specializations in contacting dendrites (*2*). Conversely, neuroligins could induce the formation of presynaptic specializations in chemically distinct axons that release glutamate or GABA (*2*). These findings confirm that the β-neurexin–neuroligin interaction can induce both excitatory and inhibitory synapses.

If the neuroligin-neurexin axis underlies the formation of excitatory and inhibitory synapses, what specifies which kind of synapse is built? There are clues that neuroligin isoforms have different functions. Whereas NL-1 and NL-3 are mainly found in excitatory synapses, NL-2 is concentrated in inhibitory synapses (*1*–*3*). NL-2 seems more potent for inducing inhibitory synapses (*1*) and it is also unique in its ability to induce clustering of gephyrin, a scaffold



Neuroligins in synapse assembly. (**Left**) At excitatory synapses, transsynaptic interactions between β-neurexin (β-NXN) and neuroligin-1 and -3 (NL-1, 3) align the presynaptic and postsynaptic membranes. In this way, the presynaptic specialization containing glutamate neurotransmitter is brought into apposition with the postsynaptic specialization containing glutamate receptors (GluRs) and the PSD-95 scaffold protein. (**Right**) In inhibitory (GABA-releasing) synapses, NL-2 may perform an analogous function by binding to β-neurexin and bringing the pre- and postsynaptic membranes into alignment.

CREDIT: NATASHA K. HUSSAIN AND PRESTON HUEY/*SCIENCE*

The authors are at the Picower Center for Learning and Memory, RIKEN-MIT Neuroscience Research Center, Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: natashah@mit.edu, msheng@mit.edu

Published by AAAS

protein found in inhibitory synapses (*2*). Unlike the direct binding between neuroligin and PSD-95, there is no known interaction between neuroligins and gephyrin, so additional undiscovered factors need to be invoked for the specification of inhibitory synapses.

Why should NL-2 behave differently from NL-1 and NL-3? An intriguing clue is that the level of PSD-95 (which binds to all three neuroligin isoforms) seems to determine whether inhibitory or excitatory synapses are made (*3*). Overexpression of PSD-95 recruited more neuroligin to glutamate synapses and reduced the relative number of GABA synapses. Conversely, suppression of PSD-95 increased inhibitory synapses at the expense of excitatory synapses (*3*). If NL-2 binds less well or is less accessible to PSD-95, then this could explain its propensity for inducing inhibitory synapses.

Whatever the precise mechanism, their double duty as inducers of either excitatory or inhibitory synapses places β-neurexins and neuroligins center stage in the control of excitation/inhibition (E/I) balance, which is critical for neuronal function (*7*). Indeed, neurons deficient in neuroligins displayed an abnormal E/I balance, with greater loss of inhibition than of excitation (*1*). Genetic mutations in NL-3 and -4 have been implicated in human mental retardation and autism (*8*–*10*). Could these illnesses be due to an E/I imbalance resulting from aberrant formation of excitatory versus inhibitory synapses? Genetic experiments in which individual β-neurexin and neuroligin isoforms are disrupted in animal models are essential to explore this idea and to confirm the conclusions reached by the intriguing in vitro studies discussed here.

References

1. B. Chih, H. Engelman, P. Scheiffele, *Science* **307**, 1324 (2005); published online 27 January 2005 (10.1126/science.1107470).
2. E. R. Graf, X. Zhang, S. X. Jin, M. W. Linhoff, A. M. Craig, *Cell* **119**, 1013 (2004).
3. O. Prange *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13915 (2004).
4. M. Missler, R. Fernandez-Chacon, T. C. Sudhof, *J. Neurochem.* **71**, 1339 (1998).
5. P. Scheiffele, J. Fan, J. Choih, R. Fetter, T. Serafini, *Cell* **101**, 657 (2000).
6. C. Dean *et al.*, *Nature Neurosci.* **6**, 708 (2003).
7. G. Liu, *Nature Neurosci.* **7**, 373 (2004).
8. D. Comoletti *et al.*, *J. Neurosci.* **24**, 4889 (2004).
9. F. Laumonnier *et al.*, *Am. J. Hum. Genet.* **74**, 552 (2004).
10. S. Jamain *et al.*, *Nature Genet.* **34**, 27 (2003).

10.1126/science.1110011

HISTORY OF SCIENCE

"How Science Survived"—Medieval Manuscripts as Fossils

Sharon Larimer Gilman and Florence Eliza Glaze

Ancient texts survived from Antiquity through the Middle Ages as manuscript copies produced by monks and professional scribes until the invention of the printing press in the 15th century A.D. Many of these texts are still in existence today, having battled an astonishing array of hazards including fire, war, theft, and neglect. But how can we calculate the percentages of texts that have survived or gone extinct and consequently the amount of knowledge that we have inherited from Antiquity and the Middle Ages? On page 1305 of this issue, Cisne (*1*) takes a unique and stimulating approach to solving this dilemma by linking the paleodemography of such texts to population dynamics. Cisne took a small number of extant medieval scientific manuscripts, such as Bede's *De Temporum Ratione* from 725 A.D., and examined them as "fossils" of early textual "populations." Applying models used by population biologists, Cisne calculated the size and age-distributions of these scientific texts. His work provides a wonderful example of the potential value of collaboration between the arts and sciences.

Enhanced online at www.sciencemag.org/cgi/content/full/307/5713/1208

The authors are in the Departments of Biology and History, Coastal Carolina University, Conway, SC 29528, USA. E-mail: sgilman@coastal.edu, fglaze@coastal.edu

Cisne applied two models from population biology to assess the survival and rate of expansion of textual populations of medieval manuscripts. These models are the Verhulst-Pearl logistic model of population growth, and the Markov birth-and-death process (where birth equals the production of new copies of texts, and death means the destruction of such texts). According to these models, the population of manuscripts increases logistically based on the standing population at any one time, taking into account birth and death rates. This works for medieval manuscripts because they are individually copied; the more there are, the faster they can reproduce, just as with a living (or a once-living) population. The rate of population growth slows as the population reaches its carrying capacity. For living organisms, this would be largely the result of increased competition for resources. For the manuscripts, it is a combination of the decline in demand for and the rate of destruction of the manuscript. The population biology models used by Cisne tend to apply best to populations of organisms with the simplest of life-histories, but the author makes a compelling case that at least some medieval manuscripts appear to have similarly simple "life-histories." The Cisne study offers a potentially useful tool to examine those histories, provided that certain modifications are applied—for the history of manuscripts isn't quite as simple as the model assumes.

To begin with, certain statements and the assumptions they betray in the Cisne report require explicit correction. It is scarcely true that, in Cisne's words, "the germ of…science barely made it through the Middle Ages." On the contrary, from the 12th through 15th centuries, science and scientific medicine constituted two of the most vigorous disciplines pursued in universities across Europe (*2*). Most Greek scientific and medical literature surviving today from the ancient world was recovered during this period, the texts of Aristotle and Galen being the best examples. In addition, new manuscripts were avidly sought and translated—both from Greek and Arabic—and these texts were commented upon in the university system that was itself a forum for discourse and disputation invented by medieval scholars (*3*). The sorry state of scientific studies at the close of the Roman Empire in the fifth century reflected Roman, not medieval, failures and shortcomings (*4*). Although the Latin language was capable of communicating scientific ideas, most Romans showed little interest in wholesale scholarly translations from Greek (*5*). The precipitous decline in Greek literacy among the Latinate population in the Western Empire by the 3rd century created a crisis in the transmission of scientific literature that was only corrected in the 12th century, after the many disruptions of the early Middle Ages had subsided and the secular school had been reborn (*6*). Cisne correctly guesses that the leap from papyrus to parchment in late Antiquity was one crucial element in the survival of texts, but there were many others (*7*). Finally, high-to-late medieval enthusiasm for science suffered at the close of the Middle Ages, when humanists of the Renaissance turned away from scientific studies (*3*, *8*). Many humanists impugned the scientific tradition

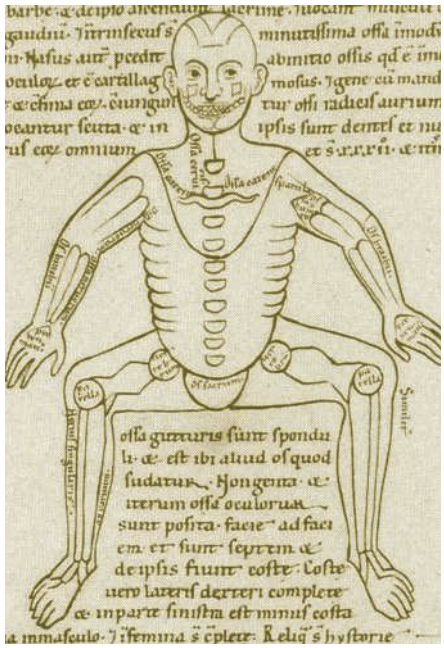
derived from Islam, and only came to embrace science in the 16th century after leading theorist-practitioners had adopted the humanists' own classicizing methods.

Having dispelled somewhat the widely held myth that the Middle Ages were a "dark age" for science, we can explore the possibilities as well as potential shortcomings of Cisne's methods. Two of the model's assumptions hold up well to scrutiny. The first is that the "population" responds immediately to changing conditions (that is, there is no time-lag to the response as might be the case with a population of living organisms); the second is that all individuals are equivalent from a quantitative perspective. Qualitatively, however, we must remember that a textual fragment can never reproduce a full, integral copy. Moreover, neither a constant environment, a closed population, nor a constant rate of text destruction (death rate) seem valid assumptions to us. Regarding the constant environment and probable half-life, there were more violations than could easily be enumerated. Many texts enjoyed a very limited circulation, owing in part to their complex theoretical language and verbosity. Others were written in regional, highly calligraphic scripts, with peculiar abbreviations and letter-forms, and therefore remained isolated or indecipherable when transported elsewhere. Death rate was inconstant, and due to many factors: intentional destruction, theft, neglect, erasure, mishandling by Renaissance printers, and modern disasters intended and accidental. Some population sites, such as monasteries, were safer than others; there are no survivors of privately owned scientific manuscripts before the 12th century. Centuries of military action in regions like the Rhineland extinguished entire collections.

Neither, as Cisne's model assumes, was the population of medieval scientific manuscripts an entity closed to immigration and emigration. Quite the opposite, in fact. From the 11th century onward, new translations made by North African and European scholars from outside populations in the Byzantine and Islamic realms were transmitted across Europe, replacing in popularity and sophistication the earlier texts from Carolingian times (late 8th to early 10th century A.D.). The earlier population of texts survived mainly in private or monastic collections, and was not considered suitable for scholastic studies. Even scientific illustrations were transmitted, sometimes "translated," as they entered the Latin West from the eastern Mediterranean. An example of this process can be seen in the anatomical text known as "the 5-Figure Series," which entered Europe in the 12th century (*9*) (see the figure).

It might behoove Cisne to refine his study by limiting and defining the focal population more carefully. Much of his evidence examines the texts that were current in the Carolingian period, which were produced before the influx of new translations after 1050 A.D. If Cisne explicitly confined his study to this population, tracing its survival in later medieval manuscripts, his assumptions would suffer fewer violations. He might also consider enumerating a broader series of texts and their representative witnesses as his sampling. Both his figures and tables on Bede's texts show remarkable promise for understanding the population's dynamic changes over time. To these, Cisne might add Pliny in his various manifestations, including the many witnesses of the *Medicina Plinii*, and the late ancient medical encyclopedists Oribasius and Alexander of Tralles. Even the early Latin translations of the Hippocratic *Aphorisms* have now been enumerated sufficiently that the population of these texts might well demonstrate the further conclusiveness of Cisne's demographic calculations and projections (*10*).

Ultimately, Cisne's ambitious efforts to analyze surviving medieval scientific manuscripts as "fossils" offers the field of manuscript scholars the opportunity to estimate and envision within a reasonable margin of error the rate of expansion of textual populations. Cisne's approach inspires us, moreover, to consider issues that might not ordinarily present themselves, including sociocultural factors like "carrying capacity" and "predation." For his efforts, and these provocative insights, manuscript scholars owe Cisne a very real debt of gratitude.



The art of translation. Populations of medieval scientific manuscripts were open to immigration and emigration. Scientific illustrations often were transmitted, or sometimes even "translated," as they entered the Latin West from the eastern Mediterranean. An example of this process can be seen in the anatomical text known as the "5-Figure Series," which entered Europe in the 12th century (*9*). (**Left**) A modern Turkish copy of a medieval Arabic anatomy figure from the 5-Figure Series, in traditional "dissection" pose. (**Right**) A 12th-century Latin Romanesque "translation" of the same figure.

CREDITS: (LEFT) COLLECTION OF ELIZA GLAZE; (RIGHT) K. SUDHOFF, *STUDIEN ZUR GESCHICHTE DER MEDIZIN*, HEFT 1, PLATE 14/BARTH, LEIPZIG, 1907

References and Notes

1. J. L. Cisne, *Science* **307**, 1305 (2005).
2. E. Sylla, M. McVaugh, Eds. *Texts and Contexts in Ancient and Medieval Science* (Brill, Leiden, 1997).
3. D. C. Lindberg, *The Beginnings of Western Science: The European Scientific Tradition in Philosophical, Religious, and Institutional Context, 600 B.C. to A.D. 1450* (Univ. of Chicago Press, Chicago, 1992).
4. W. H. Stahl, *Roman Science: Origins, Development and Influence to the Later Middle Ages* (Univ. of Wisconsin Press, Madison, 1962).
5. D. R. Langslow, *Medical Latin in the Roman Empire* (Oxford Univ. Press, Oxford, 2000).
6. F. E. Glaze, thesis, Duke University (2000); in preparation as *Medicine and Mission: The Production, Use and Circulation of Medical Books in Europe, 500–1200.*
7. M. McCormick, *Origins of the European Economy: Communications and Commerce, AD 300–900* (Cambridge Univ. Press, Cambridge, 2001).
8. R. French, *Medicine Before Science: The Business of Medicine from the Middle Ages to the Enlightenment* (Cambridge Univ. Press, Cambridge, 2003), chap. 5.
9. E. Savage-Smith, "Islamic Medical Manuscripts at the National Library of Medicine" (National Institutes of Health, Bethesda, MD, 2005); see especially www.nlm.nih.gov/hmd/arabic/p18.html and www.nlm.nih.gov/hmd/arabic/p19.html.
10. Editions and studies listed in *Bibliographie des textes médicaux latins: antiquité et haut Moyen Age*, G. Sabbah, P.-P. Corsetti, K.-D. Fischer, Eds. (Université de Saint-Etienne, Saint-Etienne, 1987).

11. We are grateful to M. H. Green, K.-D. Fischer, E. Savage-Smith, and R. Neville for their criticisms, assistance, and suggestions. All errors remain our own.

10.1126/science.1109679

The Road to Quantum Computing

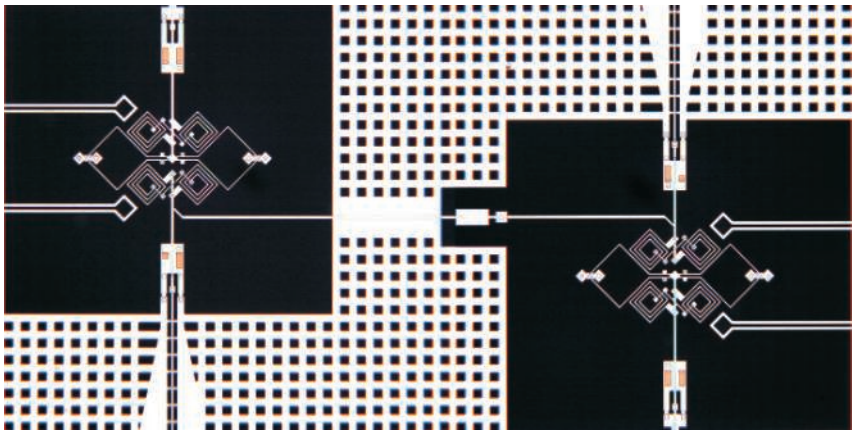
Hans Mooij

Quantum computers will be able to perform calculations that ordinary computers will never be able to, such as factorizing very large numbers (*1*). However, building a quantum computer is not easy. On page 1299 of this issue, McDermott *et al.* (*2*) report an important step toward a macroscopic version of such a computer.

The basic elements of the quantum computer are quantum bits (qubits) that, similar to ordinary bits, have two states. However, in contrast to ordinary bits, qubits can exist in combinations of states called quantum superpositions. To operate a quantum computer, one must be able to drive the qubits from any state to any other in a controlled manner. Also, one must be able to couple qubits in ways that lead to entanglement—an action-at-a-distance between quantum objects predicted by quantum theory that has been observed with photons and atoms.

Photons and atoms are obvious candidates for qubits. But for large quantum computers, it may be advantageous to use macroscopic qubits that can be fabricated with the tools of the semiconductor industry. Superconducting systems are particularly promising. Using such a system, McDermott *et al.* (*2*) report the simultaneous readout of two coupled qubits, with tools that can in principle be scaled up to a large computer. The study is an important step toward the use of artificially made objects to realize two-qubit quantum gates that perform all the needed operations for the execution of any quantum algorithm.

The qubits in question consist of a superconducting Josephson tunnel junction through which a constant current is passed. In a Josephson junction, a thin insulator is sandwiched between two superconducting metal films. The current between the two films is controlled by the phase difference between them. If the phase difference varies in time, there is a voltage between the two films. Because the junction also acts as an electrical capacitance, this voltage leads to an electrical charging energy. A phase difference across the junction increases the energy, causing an oscillation with a frequency that is determined by the parameters of the junction. In such a quantum oscillator, the amplitude of oscillation is limited to a discrete set of values.



Two fabricated quantum bits. The superconducting qubits with their control and measuring circuitry are situated at the left and right. They are coupled by a capacitor (white rectangle in the center). The qubits themselves are too small to be seen at this scale. The circuit is fabricated with standard semiconductor techniques.

The potential well in which the oscillation takes place is a metastable energy minimum, which is separated from the global minimum by an energy barrier. The height of that barrier decreases with increasing current. To detect the state of the qubit, the current through the junction is suddenly increased slightly. If the qubit is in the excited state, the junction has just enough energy to cross the barrier. The crossing of the barrier induces an observable change in a secondary circuit, which behaves almost classically and can be read out at a later stage (*3*).

The qubits described above are known as phase qubits. Other types of superconducting qubits have also been reported (*4*); charge qubits are defined by the presence or absence of a single pair of electrons, whereas in flux qubits, a persistent circulating current runs clockwise or anticlockwise. With all three types, controlled, driven transitions of individual qubits have been demonstrated. However, further improvements are needed. All solid-state quantum bits can easily couple to noise from the outside world, particularly from circuit elements that are at a higher temperature than the qubit (which operates at a temperature of about 30 mK). Quantum information is lost after a "decoherence time" of typically 0.1 to 1 μs. Qubit operations take 1 to 10 ns (1 ns = 10^{-9} s). The ratio between the two time scales must be increased to get a usable quantum computer. The measurement procedure is especially important: Strong coupling to the measuring device brings in decoherence, whereas weak coupling makes it difficult to determine the qubit state unequivocally. The quantum engineer has to optimize both the contribution to decoherence and the expected quality of readout.

The main bottleneck for the further development of superconducting qubits is the quality of the thin insulator. In today's devices, the insulator is an oxide that is either amorphous or an imperfect crystal. It therefore contains defects that move with time and that become ionized and deionized. Because transport through the insulator depends exponentially on its thickness and materials properties, displacement of just one atom over microscopic distances can have far-reaching consequences. The rate of the resulting decoherence cannot be calculated. Epitaxial or other high-quality junctions may overcome these problems. Also, different types of defects have a different influence on the various types of qubits; which qubit will be used for particular applications may therefore depend on these materials-related noise sources.

To advance from a single qubit to a full quantum computer, one first has to realize a universal two-qubit gate, preferably a quantum controlled-not (C-NOT) gate. Superconducting qubits are not there yet, but with the new results of McDermott *et al.* (*2*), they are getting close. In classical electronic circuits (with ordinary bits that can be in the 0 or the 1 state), a C-NOT gate is a conditional two-bit gate: One bit flips from 0 to 1 or vice versa if and only if the other bit is in the 1 state. The quantum C-NOT gate is particularly desirable, because all quantum

The author is at the Kavli Institute of Nanoscience, Laboratory of Applied Physics, Delft University of Technology, Delft 2628CJ, Netherlands. E-mail: j.e.mooij@tnw.tudelft.nl

Published by AAAS

algorithms can be realized, at least on paper, by combining C-NOT gates with single-qubit operations.

A C-NOT gate has previously been realized with superconducting charge qubits (*5*). However, that gate could only start from specific positions (for example, with both qubits in the 0 state), did not provide the universal operations, and cannot easily be extended to provide them. It was therefore not suitable for building a large quantum computer. The qubits of McDermott *et al*. (*2*) have the potential to yield C-NOT or similar two-qubit gates for the universal quantum computer.

The experiments of McDermott *et al*. (*2*) have several special features. Particularly attractive is the well-controlled readout. Measuring one qubit does not, as in many other schemes, disturb the whole system. Measurement of the qubit yields an up/down digitized signal, which is first stored as a current in a superconducting loop situated on the chip next to the qubit. The transfer from the qubit to the secondary circuit has to be fast, because the relaxation time of the qubit is on the order of 100 ns. The initial transfer creates only a weak disturbance and can be performed in parallel on the two qubits. The readout of the secondary circuits takes place later. The up/down signals in the secondary loops are quite robust, and measurement of one signal does not influence the other. The whole process makes it possible to perform quantum operations on the coupled system and afterwards determine the state of both qubits simultaneously.

The new experiments (*2*) do not yet demonstrate the complete two-qubit gate that is required for a quantum computer. To do so, one would have to determine the outcome after a specific set of microwave pulses is applied to the qubits, starting from different initial conditions. Also, the amplitudes of the oscillations observed in (*2*) are smaller than is acceptable for applications. Nevertheless, the study is a substantial advance. The simultaneous measurement of two qubits opens the possibility of testing the concept of entanglement for artificially fabricated quantum objects. The results prove that two-qubit operations in which quantum information is shared and manipulated are possible in solid-state qubits that can, in principle, be used to build a large quantum computer.

References

1. M. A. Nielsen, I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge Univ. Press, Cambridge, 2000).
2. R. McDermott *et al.*, *Science* **307**, 1299 (2005).
3. J. M. Martinis, S. Nam, J. Aumentado, C. Urbina, *Phys. Rev. Lett.* **89**, 117901 (2002).
4. M. H. Devoret, A. Walraff, J. M. Martinis, http://arxiv.org/abs/cond-mat/0411174.
5. T. Yamamoto, Yu. A. Pashkin, O. Astaflev, Y. Nakamura, J. S. Tsai, *Nature* **425**, 941 (2003).

10.1126/science.1109554

MICROBIOLOGY

A Pathogen Attacks While Keeping Up Defense

Staffan Normark, Christina Nilsson, Birgitta Henriques Normark

Many Gram-negative bacteria interact with human, animal, or plant hosts by injecting effector proteins into the cytosol of host cells through the so-called type III secretion system (injectisome) (*1*). The symptoms of infectious diseases, such as bubonic plague, shigellosis, salmonellosis, typhoid fever, and infantile diarrhea, largely depend on the repertoire of bacterial proteins injected by the type III secretion system, and what they do once inside eukaryotic host cells. The injectisome is remarkably well conserved among different bacterial pathogens (*2*). The multicomponent base structure spans the bacterial periplasm and is associated with both the inner and outer membranes of the bacterium (see the figure) (*3*). The filamentous needle, composed of a single protein, projects beyond the bacterial surface. For bacterial effector proteins to be translocated into host cells, the tip of the needle must make contact with the eukaryotic host cell membrane.

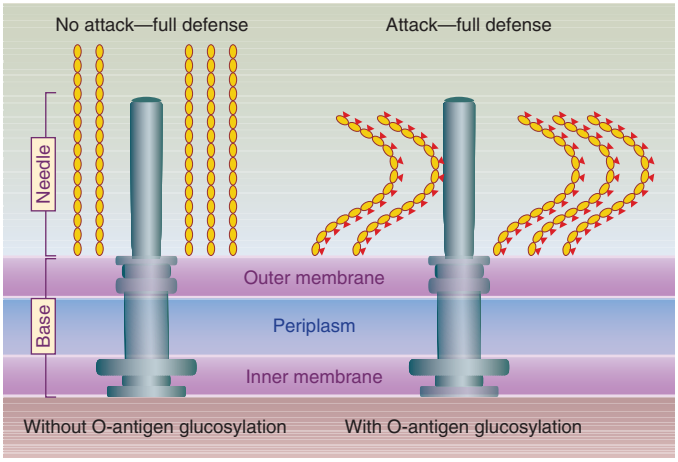
The enteric pathogen *Shigella flexneri* possesses a type III secretion system that enables invasion of the gut epithelial cells of mammalian hosts. Invasion provokes an extensive inflammatory reaction in the gut mucosa, a hallmark of shigellosis (bacterial dysentery) (*4*). To survive host inflammatory processes such as increased production of antibacterial peptides, *Shigella* is equipped with a lipolysaccharide structure in its outer membrane that contains protective repeat units of an O-antigen polysaccharide (see the figure). This O-antigen polymer extends beyond the bacterial cell surface and potentially could sterically impede the type III secretory apparatus. However, as West *et al.* (*5*) reveal on page 1313 of this issue, *Shigella* has developed an ingenious way of ensuring that its injectisome needle remains operational without compromising the ability of the O-antigen polymer to protect against host inflammatory mediators.

West *et al.* (*5*) used signature-tagged mutagenesis (*6*) to identify colonization-defective *S. flexneri* mutants in a rabbit model of shigellosis. From these attenuated mutants, they identified two genes residing on the *gtrA*, *gtrB*, *gtrV* operon of a resident bacteriophage (*7*). This operon directs the addition of a glucose residue to each O-antigen repeat unit; this glucosylation step imparts serotype specificity to different strains of *S. flexneri*. Substantially attenuated virulence was observed in glucosylation-defective *gtr* mutants of different serotypes; virulence could be restored by introducing a serotype-specific *gtr* operon. The *gtr* mutants could still produce lipopolysaccharide with the correct number of O-antigen repeats, and could withstand noxious conditions in the gut such as bile salts, complement-mediated lysis, and gut-specific antibacterial peptides. However, compared with wild-type *S. flexneri* carrying glucosylated O-antigen, the *gtr* mutants were considerably less invasive and were less able to provoke an inflammatory response, suggesting a defect in the type III secretion system. An IpaB monoclonal antibody, recognizing the tip of the needle complex, revealed a much lower exposure of the needle at the bacterial cell surface in the glucosylation-defective *gtr* mutants.

How does bacteriophage-mediated glucosylation of the O-antigen affect exposure of the type III secretion needle? West *et al.* used electron microscopy and three-dimensional molecular modeling to show that O-antigen glucosylation results in a conformational change, from a linear extended form of the repeating O-antigen polymer to a more compact structure (see the figure). This modification allows surface exposure of the protruding needle (which is roughly 60 nm long), without compromising the protective role of the O-antigen polymer.

Shigella species cause more than 1 million deaths per year from dysentery and diarrhea, and multidrug resistance of these bacteria is a rising problem. The type III secretion system of *Shigella* and other Gram-negative pathogens is an attractive target for development of new antivirulence drugs. Salicylanilides have been identified as potent inhibitors of type III secretion in the bacterium *Yersinia*, one species of which causes bubonic plague (*8*). The West

The authors are at the Microbiology and Tumor Biology Center, Karolinska Institute, Stockholm, 171 77, Sweden. E-mail: staffan.normark@stratresearch.se



Shigella's sword and shield strategy. The pathogen *Shigella flexneri* causes disease by invading gut epithelial cells, a process mediated by effector proteins delivered by its plasmid-encoded type III secretion system (gray). The host responds to replication of *Shigella* inside gut epithelial cells with an aggressive inflammatory reaction. The bacterium defends itself against antibacterial inflammatory mediators by expressing at its surface a lipopolysaccharide containing O-antigen polysaccharide repeats (yellow). These repeats extend out from the bacterial cell surface, and may impede the ability of the needle tip to contact host cells. This dilemma is solved by a resident bacteriophage that adds one glucose residue (red triangle) to each of the repeating O-antigen units. Glucosylation causes a conformational change, from a more filamentous (**left**) to a more compact O-antigen structure (**right**). The more compact structure allows exposure of the needle tip, enabling invasion of gut epithelial cells to proceed without compromising the ability of the O-antigen polymer to protect the bacterium against host inflammatory mediators.

et al. study suggests that an inhibitor of Gtr-mediated O-antigen glucosylation in *Shigella* bacteria could ameliorate the symptoms of shigellosis by preventing delivery of type III effector proteins to gut epithelial cells of the host. The GtrV protein encoded by the bacteriophage operon catalyzes the transfer of a glucosyl residue via an α1,3 linkage to rhamnose II of the O-antigen unit (*9*). GtrV is an integral membrane protein, but its periplasmic loops are assumed to be functional because O-antigen modification is thought to take place in the bacterial periplasm. Developing bacterial virulence inhibitors that target periplasmic proteins is expected to be considerably easier than developing drugs against intracellular bacterial targets.

In enteropathogenic *Escherichia coli*, one of the translocated proteins, EspA, polymerizes to form an extension to the injectisome needle (*10*), eliminating the need to modify the lipopolysaccharide for contact with its host. The length of the needle seems to vary among different pathogens. For example, the needle possessed by *Salmonella* is longer than that of *Shigella* (*2*). In *Yersinia*, the length of the needle is governed by the size of the YscP protein (*11*). There is a linear relation between needle length and the number of amino acids in YscP. Even though the mechanism for this molecular ruler remains unknown, one end of YscP may be associated with the base and the other with the growing tip of the needle. When the YscP protein is fully stretched, a signal is delivered blocking further polymerization and preventing the needle from growing any longer. As Mota *et al.* (*12*) show on page 1278 of this issue, *Yersinia* bacteria with short needles are less effective at delivering the YopP effector protein into target cells than are those with normal-length or long needles (*12*). Thus, there is a minimum needle length required for operation of the *Yersinia* injectisome. The length of the needle correlates with the length of YadA, an adhesin-forming microfilament extending about 28 nm beyond the bacterial outer membrane. Mota *et al*. demonstrate that decreasing the length of YadA in the short-needle *Yersinia* mutants suppressed the translocation defect by decreasing the distance between the needle and the host cell. As the two new studies show, needle length in both *Yersinia* and *Shigella* has evolved together with other surface structures to allow the bacterial needle tip to interact with host cell membranes. Simultaneously, elegant mechanisms have evolved in these bacterial species to ensure that the adhesive and protective functions of other filamentous surface structures are not compromised.

References

1. R. Rosqvist, K. E. Magnusson, H. Wolf-Watz, *EMBO J.* **13**, 964 (1994).
2. A. P. Tampakaki *et al.*, *Cell Microbiol.* **6**, 805 (2004).
3. T. C. Marlovits *et al.*, *Science* **306**, 1040 (2004).
4. P. Cossart, P. J. Sansonetti, *Science* **304**, 242 (2004).
5. N. P. West *et al.*, *Science* **307**, 1313 (2005).
6. M. Hensel *et al.*, *Science* **269**, 400 (1995).
7. S. Guan, D. A. Bastin, N. K. Verma, *Microbiology* **145**, 1263 (1999).
8. A. M. Kauppi *et al.*, *Chem Biol.* **10**, 241 (2003).
9. H. Korres, N. K. Verma, *J. Biol. Chem.* **279**, 22469 (2004).
10. S. J. Daniell *et al.*, *Mol. Microbiol.* **49**, 301 (2003).
11. L. Journet, C. Agrain, P. Broz, C. R. Cornelis, *Science* **302**, 1757 (2003).
12. L. J. Mota, L. Journet, I. Sorg, C. Agrain, G. R. Cornelis, *Science* **307**, 1278 (2005).

10.1126/science.1109836

RETROSPECTIVE: EVOLUTION

Ernst Mayr (1904–2005)

Jerry A. Coyne

The death of Ernst Mayr at age 100 on 3 February marks the end of a scientific era. Mayr was the last living architect of the "Modern Evolutionary Synthesis," one of the greatest intellectual achievements of 20th-century biology. His 1942 book, *Systematics and the Origin of Species*, was, along with Theodosius Dobzhansky's *Genetics and the Origin of Species* (1937), largely responsible for the seminal achievement of the Synthesis. This was the demonstration that evolutionary patterns and processes in natural populations are consistent with Darwinian natural selection, the heredity mechanisms revealed by laboratory work in genetics, and the mathematical theories of population genetics.

It is not too much of an exaggeration to call Mayr the Darwin of the 20th century. Although nobody—least of all Mayr—would claim that his stature and achievements equaled Darwin's, Mayr nevertheless solved a major problem that eluded Darwin: the origin of biodiversity. Despite the title of his greatest work, Darwin made little contribution to understanding the origin of species. Rather, he explained the origin of features *within* species. Although others contributed to explaining how new species arise, Mayr and Dobzhansky get the most credit for synthesizing and revitalizing studies of speciation.

Mayr made three major contributions to understanding biodiversity. First, along with his colleague Dobzhansky, he recognized it as an unsolved problem. Why is nature divided into discrete groups—species—rather than forming an organic continuum? And how could the gradual and continuous process of Darwinian natural

The author is in the Department of Ecology and Evolution, University of Chicago, Chicago IL 60637, USA. E-mail: j-coyne@uchicago.edu

New Perspectives on Ancient Mars

Sean C. Solomon,[1]* Oded Aharonson,[2] Jonathan M. Aurnou,[3] W. Bruce Banerdt,[4] Michael H. Carr,[5] Andrew J. Dombard,[6] Herbert V. Frey,[7] Matthew P. Golombek,[4] Steven A. Hauck II,[8] James W. Head III,[9] Bruce M. Jakosky,[10] Catherine L. Johnson,[11] Patrick J. McGovern,[12] Gregory A. Neumann,[13] Roger J. Phillips,[6] David E. Smith,[7] Maria T. Zuber[13]

Mars was most active during its first billion years. The core, mantle, and crust formed within ~50 million years of solar system formation. A magnetic dynamo in a convecting fluid core magnetized the crust, and the global field shielded a more massive early atmosphere against solar wind stripping. The Tharsis province became a focus for volcanism, deformation, and outgassing of water and carbon dioxide in quantities possibly sufficient to induce episodes of climate warming. Surficial and near-surface water contributed to regionally extensive erosion, sediment transport, and chemical alteration. Deep hydrothermal circulation accelerated crustal cooling, preserved variations in crustal thickness, and modified patterns of crustal magnetization.

Observations from spacecraft and analyses of martian meteorites suggest that the first billion years after solar system formation was a time of intense surficial and internal activity on Mars. The heavy impact bombardment of the inner solar system obscured much of the early record, but signatures of that era remain in the structure and magnetization of the martian crust and in isotope anomalies in martian meteorites. Understanding the comparative evolution of the terrestrial planets offers the potential to extract generalizations on common processes. Furthermore, the history of early Mars may offer clues to the earliest history of the larger Earth, from which even fewer remnants have been preserved. Finally, Mars has special interest as a potential habitat for past or present life, and it is during the earliest era when the boundary conditions pertinent to life were established.

Martian history has been divided into three major epochs on the basis of stratigraphic relationships and the density of impact craters (*1*–*3*). The Noachian Epoch dates terrain older than about 3.7 billion years ago (Ga) and coincides approximately with the time of heavy impact bombardment of the inner solar system (*4*). About 40% of the surface of Mars (*5*) is Noachian in age (Fig. 1A), as are all of the largest impact basins. Progressively younger epochs are the Hesperian (3.7 to ~3 Ga) and Amazonian (3 Ga to present). Of more than 30 martian meteorites—so identified on the basis of generally young crystallization ages, distinctive rock chemical and isotopic signatures, and molecular and isotopic compositions of trapped gases similar to those of the martian atmosphere (*6*)—only one (ALH84001 at ~4.5 Ga) is Noachian in age; all of the others have ages of 1.3 Ga or less and sample the Middle to Late Amazonian (*7*).

Planetary Formation and Global Differentiation

Dynamical simulations of the accretion of the terrestrial planets hint that Mars may have formed in as short an interval as several million years. Calculations of the final stages of terrestrial planet formation—the gravitational interaction over 10^8 years (*8*) of planetary embryos the size of Mercury to Mars formed in the solar nebula disk during the first 10^6 years of solar system history (*9*)—can account for planets with masses and semimajor axes similar to those of Earth and Venus, but any final body at the orbit of Mars tends to be too large (*8*, *10*). This difficulty may indicate an unusually low initial density of disk material in the vicinity of Mars's orbit (*8*). Alternatively, Mars may be a surviving embryo that escaped either accretion or ejection (*10*). This possibility would reflect the highly chaotic nature of the late-stage planetary formation process and would imply that Mars was nearly fully formed before the final stage of formation of the larger inner planets.

An important aspect of the formation process is the initial inventory of water on Mars. Dynamical and meteorite chemical arguments suggest that the environment in the solar nebula at the solar distance range of the terrestrial planets was too hot for hydration of the planetesimals; if so, the early inventory of water on the terrestrial planets must have come primarily from water-rich embryos that formed farther from the Sun and were scattered inward by interactions with the gas-giant planets (*11*). Simulations of this process yield water contents on Mars ranging from a few percent to a few tens of percent by mass of a terrestrial ocean (*10*). An alternative formation scenario involving a cooler solar nebula and hydrated planetesimals at the solar distance of Earth and Mars (*12*) would more than double the initial water inventory for Mars (*10*).

Rapid accretion of Mars could have converted into heat a sufficient quantity of kinetic energy to melt the martian interior and drive differentiation of core, mantle, and crust (*13*). Mars possesses a dense, metal-rich core (*14*) and evidence for an ancient global magnetic field produced by a core magnetic dynamo (*15*). Mars also has a silicate crust chemically different from and lower in density than the underlying silicate mantle (*16*–*19*). Isotopic anomalies in martian meteorites—all igneous rocks that are products of partial melting in the mantle or crust—provide evidence that global differentiation occurred early. Two short-lived radioactive decay systems [^{182}Hf to ^{182}W, half-life of 9 million years (My), and ^{146}Sm to ^{142}Nd, 103-My half-life] provide key observations. Because core-mantle segregation would have swept W (but not Hf) into the core, the presence of ^{182}W at levels in all martian meteorites above those for primitive chondritic meteorites (*20*, *21*) indicates that the martian core formed during the lifetime of ^{182}Hf (*22*), most likely within 10 to 15 My

[1]Department of Terrestrial Magnetism, Carnegie Institution of Washington, Washington, DC 20015, USA. [2]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125, USA. [3]Department of Earth and Space Sciences, University of California, Los Angeles, CA 90095, USA. [4]Jet Propulsion Laboratory, Pasadena, CA 91109, USA. [5]U.S. Geological Survey, Menlo Park, CA 94025, USA. [6]Department of Earth and Planetary Sciences, Washington University, St. Louis, MO 63130, USA. [7]Laboratory for Terrestrial Physics, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA. [8]Department of Geological Sciences, Case Western Reserve University, Cleveland, OH 44106, USA. [9]Department of Geological Sciences, Brown University, Providence, RI 02912, USA. [10]Laboratory for Atmospheric and Space Physics and Department of Geological Sciences, University of Colorado, Boulder, CO 80309, USA. [11]Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California at San Diego, La Jolla, CA 92093, USA. [12]Lunar and Planetary Institute, Houston, TX 77058, USA. [13]Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*To whom correspondence should be addressed. E-mail: scs@dtm.ciw.edu

of the formation time of the oldest solar system objects (*21*)—Ca-Al-rich inclusions in chondritic meteorites dated at 4.567 Ga (*23*). Because crust-mantle separation generally fractionates rare earth elements such as Sm and Nd, variations in ^{142}Nd, isotopic abundances from chondritic levels (*24*), indicate that the meteorite source regions were established while ^{146}Sm was still extant, probably within ~30 My of solar system formation (*25*, *26*). Furthermore, systematic relations among isotopes of Pb (*27*), Sr (*28*), and Os (*29*) in martian meteorites are consistent with large-scale silicate fractionation ~50 My after solar system formation and little to no remixing of crust and mantle thereafter. The 4.5-Ga age of ALH84001 (*7*) also indicates that stable crust had formed by that time. Such a rapid formation time for the crust and the mantle sources of martian meteorites suggests that these regions formed by the differentiation of a global silicate magma ocean produced by planetary accretional heating (*30*, *31*).

The Early Crust

The distribution of the present crust on Mars is constrained by the planet's topography (Fig. 1B) and gravity fields (*19*, *32*). Models that have a uniform crustal density and variable crustal thickness consistent with these data display a dichotomy in thickness (Fig. 1C): Crust in the more elevated southern hemisphere is generally thicker than crust in the northern lowlands. The boundary between these provinces varies smoothly in the vicinity of Arabia Terra (Fig. 1B), but near 90°E longitude the boundary is more steplike, in part the result of the ballistic redistribution of upper crustal material during the formation of the Hellas (*33*), Utopia, and Isidis impact basins. Determining the mean thickness of the crust requires additional considerations. Relationships between gravity and topography are sensitive to the regional thickness of the crust and suggest global average thicknesses between 30 and 80 km (*34*, *35*). These estimates, together with mass-balance calculations tied to surface abundances of U, Th, and K measured from orbit and at the Mars Pathfinder landing site (*36*) and Nd isotope systematics in martian meteorites (*37*), are consistent with an average crustal thickness of about 50 km (*35*). With this average thickness, the modal values for crustal thickness beneath the northern lowlands and southern highlands are about 35 and 60 km, respectively (*32*).

Although surface units in the northern lowlands of Mars are predominantly Hesperian to Amazonian in age (Fig. 1A), numerous partially buried impact craters and basins have been identified from their topographic signatures (*38*). The density of such features establishes that the crust of the northern lowlands and the southern uplands had formed by the Early Noachian (*38*), a result consistent with the timing of crustal formation indicated by martian meteorite data. The hemispheric dichotomy in crustal thickness, because of its global scale, must be comparable in age. Further, global thermal history models are constrained by limits on younger additions to the volume of the crust. Among models best able to reproduce Early Noachian crustal formation on Mars (*39*) are those with near-chondritic levels of interior heat production, fractionation of heat-producing elements (U, Th, and K) into the crust during crustal formation, and ductile flow laws for mantle material consistent with weakening of mantle silicates by persistent interior water (*40*).

The crustal thickness dichotomy may have been produced by spatially heterogeneous fractionation of an early magma ocean. Alternatively, crystallization of the magma ocean may have led to gravitationally unstable mantle layering, because the late-stage silicates that crystallized at shallow mantle depths were denser than earlier cumulates that crystallized near the base of the magma layer; overturn of an unstable mantle may have thickened the crust over downwelling regions and thinned the crust elsewhere (*30*). Some models of solid-state mantle convection predict a long-wavelength (harmonic degree 1) component of flow for layered viscosity structures (*41*). Such flow models might have led to thicker crust over the hemisphere dominated by upwelling and melt generation, or alternatively thinner crust over that hemisphere if flow-induced crustal thinning was more important than the effect of magmatic additions to crustal volume (*41*). An early episode of plate tectonics might have led to a crustal thickness dichotomy if crust in the two hemispheres was created at different times (*42*) or by different processes (*43*). Excavation and ballistic transport by one (*44*) or several (*45*) large impacts have also been suggested as an explanation for the crustal thickness dichotomy, but beyond the portion of the dichotomy boundary that has been influenced by the Hellas, Utopia, and Isidis basins, there is little evidence from topography (Fig. 1B) to support these suggestions.

The strongest discriminant among these hypotheses is the timing of most crustal formation. An age of ~4.5 Ga for the dichotomy rules out plate recycling (*42*, *43*) and solid-state mantle flow (*41*) on the grounds that the time scales for these processes to operate are too long. Dichotomy formation scenarios involving magma ocean evolution or impact processes are consistent with the age constraint, but the heat imparted to young crust by large impacts would tend to erase topographic relief by magmatism and crustal flow. North-south differences in crustal composition or early heat flow might be expected for some hypotheses. Differences in the infrared emission spectra have been interpreted in terms of greater average silica content for the northern hemisphere crust (*18*), a result consistent with chemical analyses of rocks, corrected for soil coating, at the Mars Pathfinder landing site (*16*). That particular chemical difference is not predicted by any of the proposed mechanisms for the formation of the crustal thickness dichotomy, however, and an alternative interpretation is that the spectra of the northern plains reflect the hydrous alteration of rocks otherwise similar to those of the southern highlands (*46*). A proxy for heat flow is the thickness of the mechanically strong outer shell or lithosphere of Mars, because the base of the lithosphere corresponds approximately to the lowest temperature at which material undergoes appreciable ductile flow over geological time scales. The thickness of the mechanical lithosphere has been estimated for southern highland regions from relationships between gravity and topography (*34*). Under the assumption that these relations were established at a time near the age of the surface units, the thickness of the mechanical lithosphere during at least a portion of the Noachian was less than 16 km, corresponding (Fig. 2) to heat flow in excess of 35 mW/m^2 (*34*). These same methods cannot be applied to address Noachian heat flow in the northern lowlands, however, because the topography has been smoothed by subsequent deposition (*47*). By the Hesperian, lithosphere thickness and heat flow in the two hemispheres were comparable (Fig. 2).

Magnetic Field History

Although Mars possesses no modern internal magnetic field of global extent, regions of the crust are strongly magnetized (*15*). These crustal magnetic anomalies likely formed when Fe-bearing minerals acquired magnetic remanence while a core dynamo was active on Mars (*15*, *48*). The largest amplitude magnetic anomalies are located in the southern uplands (Fig. 1D). Weak magnetic anomalies are resolved over portions of the northern lowlands, particularly near the boundary with the southern uplands (*49*). The youngest major impact basins on Mars (Utopia, Hellas, Isidis, and Argyre) have no magnetic anomalies resolvable from orbiting spacecraft, a result attributed to the demagnetizing effects of the shock and heating that accompany basin formation (*50*). This explanation requires that the core dynamo had ceased to be active by the time these basins formed and cooled (*15*), in contrast to the case for older and more topographically subdued basins that display prominent magnetic anomalies (*51*).

A core dynamo on Mars requires that there be sufficient energy to drive core convection, energy most likely supplied by some

combination of cooling of the core and growth of a solid inner core (*13*). The core formation process likely imparted sufficient internal energy to superheat the molten core by several hundred degrees and to initiate a dynamo in the earliest Noachian (*13*, *52*). Arguments favoring a dynamo that was active in the Early Noachian but had ceased by Late Noachian or Early Hesperian include the pronounced concentration of regions of high magnetization in Noachian terrain (*15*, *48*, *49*), a lack of correlation of magnetic anomalies with Hesperian or younger volcanic units or impact structures, and the observation of magnetic carriers within carbonates formed 3.9 Ga or earlier in martian meteorite ALH84001 (*53*). Several hypotheses can account for termination of the dynamo action, including loss of core heat (*52*), rapid solidification of most of the core (*13*), and a change in the efficiency of mantle convection to transport heat lost from the core (*54*). That some portion of the core is presently fluid, on the basis of the response of Mars to solar tides (*55*), provides a limit to core thermal history models but does not distinguish among these proposals.

The processes responsible for the observed magnetization of the crust are uncertain. The amplitude of the strongest magnetic anomalies implies that the specific magnetization in such areas is comparable to or greater than the highest common values in Earth's crust and that the vertical extent of coherent magnetization is several tens of kilometers (*48*, *56*, *57*). The elongated shape of some magnetic anomalies has led to the suggestion that magnetization in such regions was acquired during crustal spreading (*48*) or during the cooling of long dike swarms (*58*). Hydrothermal metamorphism has been postulated as a source of magnetization (*59*), and hydrothermal activity has been invoked to account for an apparent spatial correlation (*60*) between high magnetization and water-carved valley networks (*61*).

If large portions of the Early Noachian crust of the northern hemisphere were once magnetized as strongly as the areas of large-amplitude anomalies in the southern uplands, then some process must have weakened the magnetization. Burial by sediments and postdynamo lavas can reduce the anomaly magnitudes, but the preservation of topographic signatures of Early Noachian impact features limits the thickness of superposed younger sedimentary and volcanic material to 1 to 2 km, except in the central regions of the largest impact basins (*38*). Given such a limit, as long as the thickness of magnetized material beneath the northern lowlands was originally comparable to that in the areas of strongest magnetic anomalies, Noachian magnetization at wavelengths seen in the southern highlands should be detectable by orbiting spacecraft. Reheating by postdynamo volcanism and intrusion is probably also only a small contributor, because cooling of thin flows occurs rapidly without deep penetration of heat, and intrusions are unlikely to have demagnetized the crust on vertical scales comparable to the thickness of coherent magnetization (*62*).

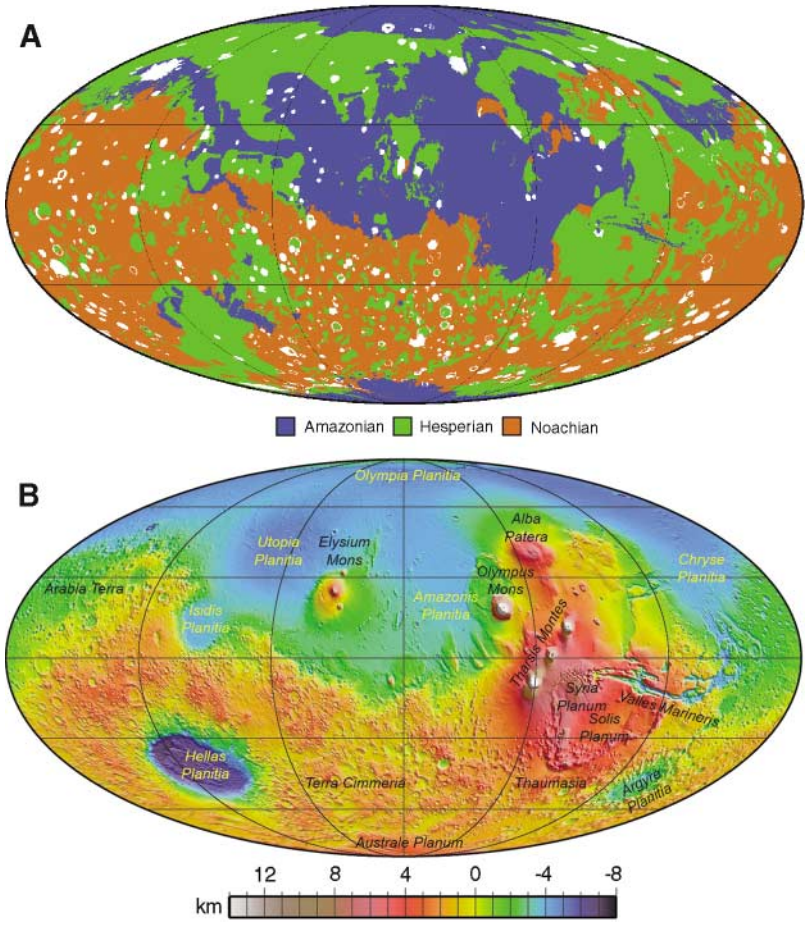
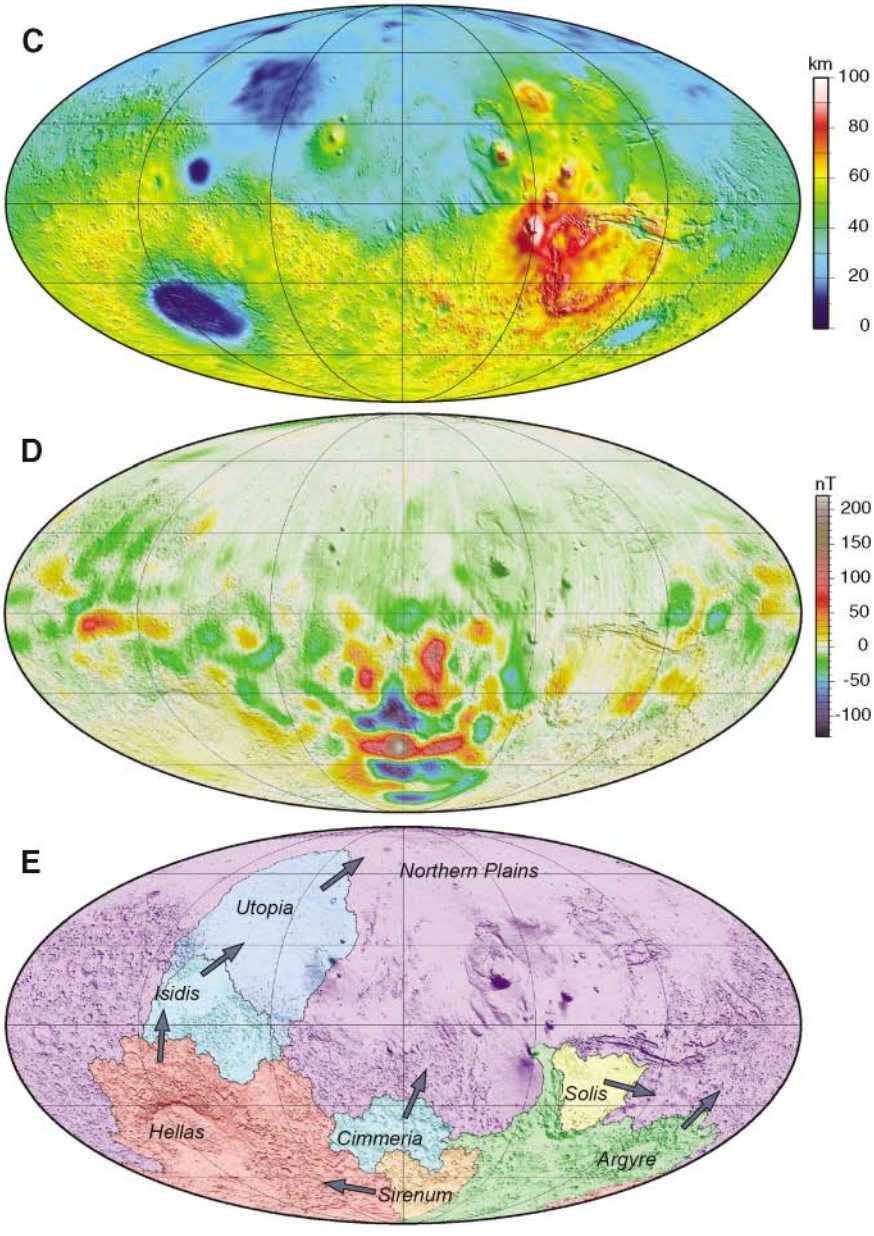


Fig. 1. (**A**) Map of martian surface units grouped on the basis of age (*1*–*3*). Units transitional between Noachian and Hesperian and between Hesperian and Amazonian have been included in the younger of the two epochs. Areas in white are impact craters and their ejecta deposits. (**B**) Topographic map of Mars (*32*, *33*) with major regions noted. On facing page: (**C**) Crustal thickness on Mars (*32*) for a density contrast at the crust-mantle boundary of 600 kg/m^3. (**D**) Radial component of the magnetic field arising from crustal magnetic anomalies on Mars (*105*) at an altitude of 400 ± 30 km. (**E**) The eight largest closed drainage basins on Mars, deduced from current topography and distinguished by color (*102*); boundaries are equivalent to continental divides on Earth. Arrows denote overflow points for each basin. All maps are in Mollweide projection, with 180° longitude at the central meridian.

Formation of Tharsis

By the end of the Middle Noachian at ~3.8 Ga (*4*), much of the magmatism and associated deformation of the crust and lithosphere on Mars had become focused within the Tharsis province, a generally elevated (Fig. 1B) area occupying ~25% of the surface (*63*, *64*). Tharsis likely originated as a center of concentrated activity after the establishment of the hemispheric difference in crustal thickness (Fig. 1C), on the basis of the age and global extent of the dichotomy and the indication from magnetic anomalies (Fig. 1D) that Tharsis volcanic units were emplaced on older magnetized crust (*62*). Patterns of uplift and volcanism in Syria Planum (Fig. 3A) and Thaumasia (*65*), layered deposits at least 8 km thick in the walls of Valles Marineris interpreted to be solidified magma (*66*), and evidence from gravity-topography relations for widespread crustal underplating (*34*) in the region are consistent with heating and melt generation across a broad zone of the mantle beneath Tharsis. The erupted and intruded magmatic material grew to exert a large downward load on the lithosphere. Comparison of models of lithospheric strain with observed deformational features suggests that by Late Noachian, the lithospheric load at Tharsis



was similar in scale and magnitude to that seen at present (*67*). The Tharsis region nonetheless remained a site of recurring volcanism for most of martian history, contains volcanic deposits with some of the youngest surface ages on the planet (*4*), and is one of two likely source areas for the Amazonian-age martian meteorites (*16*).

One or more long-lived plumes of hot upwelling mantle may account for the formation of the Tharsis province and its extended magmatic and tectonic activity. The form and vigor of mantle convection on Mars, however, depend on the extent of compositional layering, the viscosity of mantle material, and the distribution of internal heat sources, characteristics that are not well known. Early heat loss from the core during the lifetime of the martian dynamo may have rendered unstable the lower thermal boundary layer of the martian mantle, favoring the formation of localized upwellings of hot material (*13*). The base of the martian mantle is at a pressure comparable to the base of Earth's upper mantle; if the martian mantle extends to the depth at which wadsleyite [$(Mg,Fe)_2SiO_4$] transforms to perovskite [$(Mg,Fe)SiO_3$] plus magnesiowüstite [(Mg,Fe)O], that endothermic transformation can localize convective transport from the lower thermal boundary layer into one or two major upwellings (*68*). The time necessary for such localization, however, is too long (*68*) to account for the formation of a single dominant volcanic province by the Middle Noachian. If the thicker crust of the southern highlands acts as an insulating layer to the convecting mantle and if the mantle is compositionally stratified into an upper and lower layer (*30*, *31*), then laboratory analog experiments suggest that broad upwelling of the lower mantle will tend to be centered beneath the insulating layer (*69*). Furthermore, narrow and hot upper mantle plumes will tend to rise from the mantle compositional boundary above the broad zone of lower mantle upwelling (*69*). This hypothesis does not account, however, for the fact that Tharsis is centered at the dichotomy boundary between crustal provinces. Focused mantle upwelling is not a necessary condition for Tharsis; if the province originated over an upper mantle warmer than average, the lithosphere was kept thinner than average at any given time by sustained magmatism, and magma transport was facilitated by deformation localized by the thin lithosphere (*70*). Anomalously high temperature and melt production rates in the mantle beneath the Tharsis province may have been long-lived remnants of heat deposited by a large impact (*71*), but if so, any signature of that particular impact has been overprinted by volcanism.

Water and Early Climate

Mars has surficial and near-surface water at present, including water ice in both polar caps (*72*, *73*) and water ice in the shallow regolith identified from gamma-ray and neutron spectrometry (*74*, *75*). There are also morphologic indicators of subsurface water ice (*76*) and gullies in crater walls interpreted as products of recent groundwater seepage and surface runoff (*77*). An atmospheric D/H ratio five times the terrestrial value (*78*) indicates that Mars has lost a substantial fraction of its water to space by a mass-dependent mechanism, most likely involving stripping of O by the solar wind and thermal escape of H (*60*). The comparatively unfractionated N and noble gases in samples of martian atmosphere trapped in ALH84001 at ~4 Ga (*79*) support the inference that the magnetic field shielded most of the atmosphere from solar wind stripping during the Noachian (*80*), but large impacts would have removed atmospheric material including water without isotopic fractionation during the heavy impact bombardment (*80*, *81*). A larger inventory of mobile near-surface water during the Noachian than at present may be inferred from these observations.

There is evidence for pervasive interaction of liquid water with the martian surface during the Noachian. Networks of tributaries formed by groundwater sapping or precipitation-derived runoff dissect broad segments of the highlands (Fig. 3B), and most of these valley networks date from the Noachian (*82*, *83*). The flow directions of Late Noachian valley networks indicate that the majority of these

features formed after most of the Tharsis rise had been emplaced (*64*). Large areas of the highlands display morphologies (Fig. 3B) indicative of removal of 1 km or more of material during the Noachian, most likely by fluvial erosion (*84*, *85*), and observations at the Mars Pathfinder landing site limit mean erosion rates in post-Noachian epochs to values (0.01 to 0.1 nm/yr) three to six orders of magnitude lower than Noachian rates (*86*). Hydrothermally altered minerals have been identified in the highlands (*87*), but the widespread occurrence of unaltered minerals (*88*) argues that surficial water-rock interactions were not ubiquitous in the ancient crust.

The largest episodes of erosion of crustal material are recorded in outflow channels thought to be formed by large flood events (*76*). Although most identified outflow channels are Hesperian and younger in age, examples of large Noachian flood events are recognized (*89*). Material removed from the highlands by valley systems and flood events was deposited in the northern lowlands and the floors of large impact basins (*47*). Deposition in the northern plains, by some mix of sedimentary and volcanic processes, occurred at high rates during the Noachian, on the grounds that younger impact basins are superposed on the material that filled in the several-kilometer-deep Utopia basin. The delivery of large volumes of waterborne sediments to lower elevations implies that there were standing bodies of water on Mars for at least limited time periods. The Mars Exploration Rover Opportunity documented sedimentary rocks with compositions and textures indicative of episodic inundation by shallow surface water (*90*). Arguments have been advanced for a long-lived, ice-covered Noachian ocean on Mars (*91*), but testing this hypothesis is difficult because of the extensive impact and erosional degradation of the surface that occurred during that epoch and the subsequent resurfacing of lowland areas (*92*).

The evidence for liquid water at or near the surface and the high rates of erosion may indicate a generally warmer climate in the Noachian than during the later epochs that postdated the loss of much of the early atmosphere. Alternatively, warm periods may have been infrequent and short-lived. Large impacts may have melted or evaporated subsurface ice and injected water into the atmosphere that rained out over periods of years to decades, forming valley systems and recharging subsurface aquifers during those intervals (*93*). Release to the atmosphere of magmatic water and other volatiles during major volcanic episodes could have led to similar warming intervals (*64*). There has been no agreement on the details of the early climate, however, or whether there were prolonged periods when liquid water was stable at the surface (*60*).

Some Interconnections

Many of the early global-scale processes on Mars were interconnected, and an important linking element was water. During martian differentiation, the mantle retained substantial quantities of water contributed during growth of the planetary embryos from which Mars was assembled (*10*–*12*). Mantle convective thermal history models provide a better fit to early crustal formation if the mantle viscosity matches that appropriate for mantle mineral assemblages containing water (*39*). Estimates of the pre-eruptive water contents of the magmas parental to young martian meteorites as high as 1 to 2% by weight are consistent with mantle water contents of several hundredths of a percent (*94*), indicating that partial melting beneath long-lived volcanic centers such as those of the Tharsis province has not fully degassed the mantle. The low viscosity of mantle material containing water would have favored efficient convection of the lowermost mantle during the Noachian. Such convection would have cooled the core during the era of the active dynamo and would have facilitated the formation of zones of upwelling from an unstable mantle lower thermal boundary layer that might have contributed to the formation of Tharsis and other smaller volcanic centers.

Water may have helped to preserve variations in the thickness of the crust. Gravity-topography relations for the southern highlands indicate that in the Early to Middle Noachian only the upper crust displayed long-term strength (*34*). Stress differences arising from heterogeneities in crustal thickness would have driven lateral flow of lower crustal material (*19*) until the temperature-dependent ductile strength of the crust was lowered by cooling to a level sufficient to arrest such flow. The rate of cooling of the crust can be sensitive to the depth and vigor of hydrothermal circulation (*95*). Although the depth of hydrothermal circulation on Mars is not known, in young oceanic lithosphere on Earth the circulation of seawater through faults and fractures can extend to depths of at least 10 km (*96*, *97*). The confining pressure at a depth of 10 km on Earth, about 300 MPa, would be reached at a depth of about 25 km on Mars, and for such a depth, the effect of hydrothermal circulation can lengthen by many orders of magnitude the time scale for relaxation of crustal thickness variations (*95*). During the period that flow of the lower crust proceeded at high rates, the longest wavelength variations in crustal thickness would have persisted for the longest durations (*19*). The north-to-south variation in crustal thickness and topography preserved today may have been in part the result of hydrothermal cooling that arrested lower crustal flow after relaxation of many of the shorter wavelength variations.

Water and other volatiles may have been delivered to the atmosphere of Mars during major eruption events in sufficient quantity to have modified the atmosphere and the climate (*98*). The formation of Tharsis in particular may have exerted a strong perturbation to the near-surface water budget and climate during the Noachian (*64*). From modern topography and gravity, the volume of magmatic material added to Tharsis, most of it during the Noachian, has been estimated at $\sim 3 \times 10^8$ km^3 (*64*). For water contents in the parental magmas as high as 2% by weight

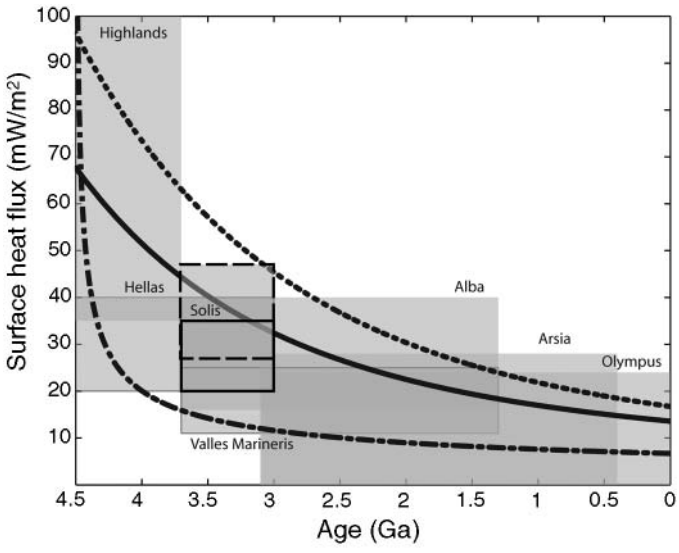


Fig. 2. Estimates of martian heat flux versus time (*34*, *106*). Most boxes denote ranges in values obtained from estimates of lithosphere thickness derived from gravity-topography relations and assigned to the epoch corresponding to the surface units in the region (*34*). One set of estimates (dashed box), derived from mechanical models for the formation of Hesperian wrinkle ridges in the northern plains (*106*), overlaps those from gravity and topography obtained for the southern uplands during that epoch (*34*). Curves (*106*) show heat flux versus time for two heat-production models and a lithospheric cooling model (dash dots).

(*94*), the water released by the emplacement of Tharsis would be equivalent to a global layer 120 m thick (*64*). For a CO_2 content in the parental magmas comparable to those in Hawaiian basalts (0.65% by weight), the construction of Tharsis could produce the integrated equivalent of 1.5 bar (0.15 MPa) of atmospheric CO_2 (*64*). Such an atmosphere could have warmed to a level whereby liquid water would be stable as a surface phase, although climate models are highly sensitive to assumptions regarding cloud formation and other processes (*99*).

Water also may have aided in determining the pattern of magnetic anomalies now seen on Mars, specifically the paucity of resolved magnetic anomalies in the lowest portions of the northern lowlands and the best preserved impact basins. Hydrothermal circulation along deep faults may have altered the magnetic carriers in the crust, by processes similar to the oxidation and reduction in magnetization that accompany circulation of seawater in terrestrial oceanic crust (*100*). Deep circulation of water would have been driven by deep-seated heat (*101*) associated with the formation of major impact structures and with local and regional magmatism. Hydrothermal alteration would have lessened the specific magnetization of crustal rocks and changed the horizontal scale and possibly the direction of coherent magnetization. A depth of water circulation of ~25 km would have been sufficient to affect most of the thickness of magnetized material seen elsewhere on Mars (*48*, *56*, *57*). The dominant horizontal scale of coherent magnetization after alteration, if comparable to the depth of circulation, would be too small for magnetic anomalies to be detectable from an orbiting spacecraft. Such anomalies would be detectable at the surface or at low altitudes.

The preferred sites for hydrothermal alteration would have been the lowest lying areas of major drainage basins (Fig. 1E), on the grounds that such areas would have been persistent traps for crustal fluids following episodic flooding and lesser discharge events. The central regions of the five drainage basins (*102*) with the largest areas and greatest topographic relief (Fig. 1E) are generally areas of weak to unresolved magnetic anomalies (Fig. 1D). Furthermore, magnetic anomalies should be attenuated or unresolvable from orbit within topographically well-preserved impact basins, even if the basins formed while a core dynamo was still active, because of the tendency of such basins to collect water and the sustained deep circulation that would have been driven by impact heating of the subsurface. That older basins preserve substantial volumes of crust remagnetized during loss of impact heat and any later magmatism (*51*) may be because the initial topographic relief was relaxed by crustal and mantle flow, lessening the likelihood that substantial volumes of water collected in the basin subsurface. The conditions favoring an active core dynamo may therefore have postdated the youngest impact basins and persisted until at least the end of the Late Noachian.

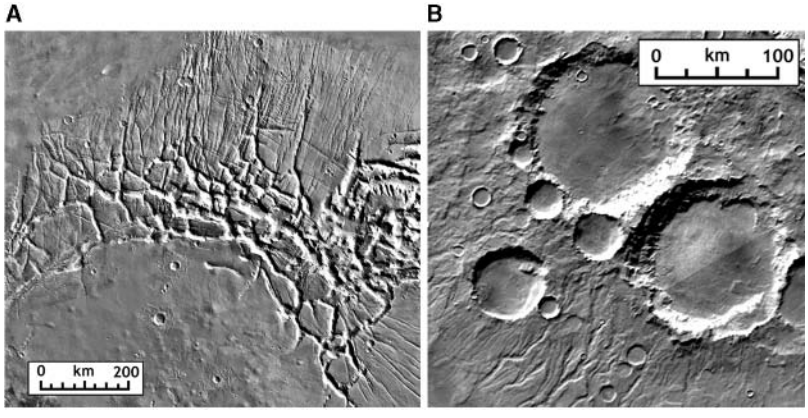


Fig. 3. Noachian-age geologic features and terrains on Mars. (**A**) Much of the volcanic construction and associated deformation in the Tharsis province during the Noachian was centered on Syria Planum (Fig. 1B). The polygonally fractured western Noctis Labyrinthus marks the present summit region (middle); ancient radially fractured plains, some of Noachian age (*1*), extend to the north, where lava flows and small shields from the younger Tharsis Montes region embay and are superposed on the ancient fractured deposits. Image was created by Viking Mars Digital Image Mosaic (MDIM) over Mars Orbiting Laser Altimeter shaded relief. (**B**) Noachian impact craters are generally highly degraded. Examples in southeastern Terra Cimmeria (47.3°S, 166.6°E) are generally rimless, partially filled craters surrounded by relatively smooth intercrater areas mapped as Noachian plains (*2*). A Noachian valley network drains downslope to the lower left. Image was created by Viking MDIM.

The depth and vigor of hydrothermal circulation within the crust may have influenced the extent of organic chemical synthesis and zones of potential habitability during the Noachian Epoch. Oxidation-reduction reactions at water-mineral interfaces in hydrothermal environments can fix carbon into organic compounds (*103*) and could have provided energy for metabolic processes (*104*). What is emerging from a better understanding of the early history of Mars is that the Noachian was a time when hydrothermal systems capable of fostering organic synthesis and sustaining any life forms that may have occupied such niches were widespread and may have extended at least halfway to the base of the crust.

References and Notes

1. D. H. Scott, K. L. Tanaka, *U.S. Geol. Surv. Misc. Geol. Invest. Map I-1802-A* (1986).
2. R. Greeley, J. E. Guest, *U.S. Geol. Surv. Misc. Geol. Invest. Map I-1802-B* (1987).
3. K. L. Tanaka, D. H. Scott, *U.S. Geol. Surv. Misc. Geol. Invest. Map I-1802-C* (1987).
4. W. K. Hartmann, G. Neukum, *Space Sci. Rev.* **96**, 165 (2001).
5. K. L. Tanaka, N. K. Isbell, D. H. Scott, R. Greeley, J. E. Guest, in *Proceedings of the 18th Lunar and Planetary Science Conference*, G. Ryder, Ed. (Cambridge Univ. Press, Cambridge, 1988), pp. 665–678.
6. H. Y. McSween Jr., *Meteoritics* **29**, 757 (1994).
7. L. E. Nyquist *et al.*, *Space Sci. Rev.* **96**, 105 (2001).
8. J. Chambers, *Icarus* **152**, 205 (2001).
9. S. J. Kortenkamp, E. Kokubo, S. J. Weidenschilling, in *Origin of the Earth and Moon*, R. M. Canup, K. Righter, Eds. (Univ. of Arizona Press, Tucson, AZ, 2000), pp. 85–100.
10. J. I. Lunine, J. Chambers, A. Morbidelli, L. A. Leshin, *Icarus* **165**, 1 (2003).
11. A. Morbidelli *et al.*, *Meteoritics Planet. Sci.* **35**, 1309 (2000).
12. M. J. Drake, K. Righter, *Nature* **416**, 39 (2002).
13. D. J. Stevenson, *Nature* **412**, 214 (2001).
14. W. M. Folkner, C. F. Yoder, D.-N. Yuan, E. M. Standish, R. A. Preston, *Science* **278**, 1749 (1997).
15. M. H. Acuña *et al.*, *Science* **284**, 790 (1999).
16. H. Y. McSween Jr., T. L. Grove, M. B. Wyatt, *J. Geophys. Res.* **108**, 5135, 10.1029/2003JE002175 (2003).
17. H. Wänke, J. Brückner, G. Dreibus, R. Rieder, I. Ryabchikov, *Space Sci. Rev.* **96**, 317 (2001).
18. J. L. Bandfield, V. E. Hamilton, P. R. Christensen, *Science* **287**, 1626 (2000).
19. M. T. Zuber *et al.*, *Science* **287**, 1788 (2000).
20. Q. Yin *et al.*, *Nature* **418**, 949 (2002).
21. T. Kleine, C. Münker, K. Mezger, H. Palme, *Nature* **418**, 952 (2002).
22. D.-C. Lee, A. N. Halliday, *Nature* **388**, 854 (1997).
23. Y. Amelin, A. N. Krot, I. D. Hutcheon, A. A. Ulyanov, *Science* **297**, 1678 (2002).
24. C. L. Harper Jr., L. E. Nyquist, B. Bansal, H. Wiesmann, C.-Y. Shih, *Science* **267**, 213 (1995).
25. L. E. Borg, L. E. Nyquist, H. Wiesmann, C.-Y. Shih, Y. Reese, *Geochim. Cosmochim. Acta* **67**, 3519 (2003).
26. C. N. Foley *et al.*, *Geochim. Cosmochim. Acta*, in preparation.
27. J. H. Chen, G. J. Wasserburg, *Geochim. Cosmochim. Acta* **50**, 955 (1986).
28. L. E. Borg, L. E. Nyquist, L. A. Taylor, H. Wiesmann, C.-Y. Shih, *Geochim. Cosmochim. Acta* **61**, 4915 (1997).
29. A. D. Brandon, R. J. Walker, J. W. Morgan, G. G. Goles, *Geochim. Cosmochim. Acta* **64**, 4083 (2000).
30. P. C. Hess, E. M. Parmentier, *Lunar Planet. Sci.* **32**, 1319 (abstr.) (2001).
31. L. Elkins-Tanton, E. M. Parmentier, P. C. Hess, *Meteoritics Planet. Sci.* **38**, 1753 (2003).
32. G. A. Neumann *et al.*, *J. Geophys. Res.* **109**, E08002, 10.1029/2004JE002262 (2004).
33. D. E. Smith *et al.*, *Science* **284**, 1495 (1999).
34. P. J. McGovern *et al.*, *J. Geophys. Res.* **107**, 5136, 10.1029/2002JE001854 (2002).
35. M. A. Wieczorek, M. T. Zuber, *J. Geophys. Res.* **109**, E01009, 10.1029/2003JE002153 (2004).
36. S. M. McLennan, *Geophys. Res. Lett.* **28**, 4019 (2001).
37. M. D. Norman, *Meteoritics Planet. Sci.* **34**, 439 (1999).

38. H. V. Frey, J. H. Roark, K. M. Shockey, E. L. Frey, S. E. H. Sakimoto, *Geophys. Res. Lett.* **29**, 1384, 10.1029/2001GL013832 (2002).
39. S. A. Hauck II, R. J. Phillips, *J. Geophys. Res.* **107**, 5052, 10.1029/2001JE001801 (2002).
40. S.-i. Karato, P. Wu, *Science* **260**, 771 (1993).
41. S. Zhong, M. T. Zuber, *Earth Planet. Sci. Lett.* **189**, 75 (2001).
42. N. H. Sleep, *J. Geophys. Res.* **99**, 5639 (1994).
43. A. Lenardic, F. Nimmo, L. Moresi, *J. Geophys. Res.* **109**, E02003, 10.1029/2003JE002172 (2004).
44. D. E. Wilhelms, S. W. Squyres, *Nature* **309**, 138 (1984).
45. H. V. Frey, R. A. Schultz, *Geophys. Res. Lett.* **15**, 229 (1988).
46. M. B. Wyatt, H. Y. McSween Jr., *Nature* **417**, 263 (2002).
47. K. L. Tanaka, J. A. Skinner Jr., T. M. Hare, T. Joyal, A. Wenker, *J. Geophys. Res.* **108**, 8043, 10.1029/2002JE001908 (2003).
48. J. E. P. Connerney *et al.*, *Science* **284**, 794 (1999).
49. J. E. P. Connerney, M. H. Acuña, N. F. Ness, D. L. Mitchell, R. P. Lin, *Lunar Planet. Sci.* **35**, 1114 (abstr.) (2004).
50. L. L. Hood, N. C. Richmond, E. Pierazzo, P. Rochette, *Geophys. Res. Lett.* **30**, 1281, 10.1029/2002GL016657 (2003).
51. H. V. Frey, *Lunar Planet. Sci.* **34**, 1838 (abstr.) (2003).
52. J.-P. Williams, F. Nimmo, *Geology* **32**, 97 (2004).
53. B. P. Weiss *et al.*, *Earth Planet. Sci. Lett.* **201**, 449 (2002).
54. F. Nimmo, D. J. Stevenson, *J. Geophys. Res.* **105**, 11969 (2000).
55. C. F. Yoder, A. S. Konopliv, D. N. Yuan, E. M. Standish, W. M. Folkner, *Science* **300**, 299 (2003).
56. F. Nimmo, M. S. Gilmore, *J. Geophys. Res.* **106**, 12315 (2001).
57. R. L. Parker, *J. Geophys. Res.* **108**, 5006, 10.1029/2001JE001760 (2003).
58. F. Nimmo, *Geology* **28**, 391 (2000).
59. E. R. D. Scott, M. Fuller, *Earth Planet. Sci. Lett.* **220**, 83 (2004).
60. B. M. Jakosky, R. J. Phillips, *Nature* **412**, 237 (2001).
61. K. P. Harrison, R. E. Grimm, *J. Geophys. Res.* **107**, 5025, 10.1029/2001JE001616 (2002).
62. C. L. Johnson, R. J. Phillips, *Earth Planet. Sci. Lett.* **230**, 241 (2005).
63. R. C. Anderson *et al.*, *J. Geophys. Res.* **106**, 20563 (2001).
64. R. J. Phillips *et al.*, *Science* **291**, 2587 (2001).
65. J. M. Dohm, K. L. Tanaka, *Planet. Space Sci.* **47**, 411 (1999).
66. A. S. McEwen, M. C. Malin, M. H. Carr, *Nature* **397**, 584 (1999).
67. W. B. Banerdt, M. P. Golombek, *Lunar Planet. Sci.* **31**, 2038 (abstr.) (2000).
68. H. Harder, U. R. Christensen, *Nature* **380**, 507 (1996).
69. M. J. Wenzel, M. Manga, A. M. Jellinek, *Geophys. Res. Lett.* **31**, L04702, 10.1029/2003GL019306 (2004).
70. S. C. Solomon, J. W. Head, *J. Geophys. Res.* **87**, 9755 (1982).
71. C. C. Reese, V. S. Solomatov, J. R. Baumgardner, *J. Geophys. Res.* **107**, 5082, 10.1029/2000JE001474 (2002).
72. T. N. Titus, H. H. Kieffer, P. R. Christensen, *Science* **299**, 1048 (2003).
73. J.-P. Bibring *et al.*, *Nature* **428**, 627 (2004).
74. W. V. Boynton *et al.*, *Science* **297**, 81 (2002).
75. W. C. Feldman *et al.*, *Science* **297**, 75 (2002).
76. V. R. Baker, *Nature* **412**, 228 (2001).
77. M. C. Malin, K. S. Edgett, *Science* **288**, 2330 (2000).
78. T. Owen, J. P. Maillard, C. de Bergh, B. L. Lutz, *Science* **240**, 1767 (1988).
79. K. J. Mathew, K. Marti, *J. Geophys. Res.* **106**, 1401 (2001).
80. D. A. Brain, B. M. Jakosky, *J. Geophys. Res.* **103**, 22689 (1998).
81. H. J. Melosh, A. Vickery, *Nature* **338**, 487 (1989).
82. M. H. Carr, *J. Geophys. Res.* **100**, 7479 (1995).
83. B. M. Hynek, R. J. Phillips, *Geology* **31**, 757 (2003).
84. R. A. Craddock, T. A. Maxwell, *J. Geophys. Res.* **98**, 3453 (1993).
85. B. M. Hynek, R. J. Phillips, *Geology* **29**, 407 (2001).
86. M. P. Golombek, N. T. Bridges, *J. Geophys. Res.* **105**, 1841 (2000).
87. P. R. Christensen *et al.*, *J. Geophys. Res.* **105**, 9623 (2000).
88. V. E. Hamilton, P. R. Christensen, H. Y. McSween Jr., J. L. Bandfield, *Meteoritics Planet. Sci.* **38**, 871 (2003).
89. A. G. Fairén *et al.*, *Icarus* **165**, 53 (2003).
90. S. W. Squyres *et al.*, *Science* **306**, 1709 (2004).
91. S. M. Clifford, T. J. Parker, *Icarus* **154**, 40 (2001).
92. M. H. Carr, J. W. Head III, *J. Geophys. Res.* **108**, 5042, 10.1029/2002JE001963 (2003).
93. T. L. Segura, O. B. Toon, A. Colaprete, K. Zahnle, *Science* **298**, 1977 (2002).
94. H. Y. McSween Jr. *et al.*, *Nature* **409**, 487 (2001).
95. E. M. Parmentier, M. T. Zuber, *Lunar Planet. Sci.* **33**, 1737 (abstr.) (2002).
96. R. T. Gregory, H. P. Taylor Jr., *J. Geophys. Res.* **86**, 2737 (1981).
97. S. C. Solomon, D. R. Toomey, *Annu. Rev. Earth Planet. Sci.* **20**, 329 (1992).
98. M. A. Bullock, D. H. Grinspoon, R. J. Phillips, *Eos* **82** (suppl.), F708 (2001).
99. M. A. Mischna, J. F. Kasting, A. Pavlov, R. Freedman, *Icarus* **145**, 546 (2000).
100. H. P. Johnson, J. E. Pariso, *J. Geophys. Res.* **98**, 435 (1993).
101. B. J. Travis, N. D. Rosenberg, J. N. Cuzzi, *J. Geophys. Res.* **108**, 8040, 10.1029/2002JE001877 (2003).
102. W. B. Banerdt, A. Vidal, *Lunar Planet. Sci.* **32**, 1488 (abstr.) (2001).
103. E. L. Shock, *J. Geophys. Res.* **102**, 23687 (1997).
104. E. S. Varnes, B. M. Jakosky, T. M. McCollom, *Astrobiology* **3**, 407 (2003).
105. J. E. P. Connerney *et al.*, *Geophys. Res. Lett.* **28**, 4015 (2001).
106. L. G. J. Montesi, M. T. Zuber, *J. Geophys. Res.* **108**, 5048, 10.1029/2002JE001974 (2003).

107. This paper began as a series of scientific discussions held at meetings of the Mars Orbiting Laser Altimeter team on the Mars Global Surveyor mission. We thank J. Dickson, L. Montesi, and J. Roark for assistance with figure preparation. Support for this paper has been provided by NASA, through the Mars Exploration Program, the Planetary Geology and Geophysics Program, and the NASA Astrobiology Institute.

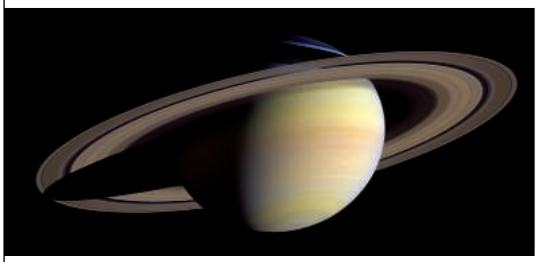
10.1126/science.1101812

INTRODUCTION

Cassini Drops In

After 23 years of limited Earth-based observations, a spacecraft has returned to Saturn and, unlike the Pioneer 11, Voyager 1, and Voyager 2 flybys from 1979 to 1981, Cassini intends to stay awhile, at least 4 years or 74 orbits, to collect information about the planet's atmosphere, its elusive magnetosphere and ionosphere, the rings, and the moons. This special issue covers the observations made during approach (December 2003 to June 2004), during the ring plane crossing and orbit insertion (1 July 2004), and the first few orbits (July through December 2004).

The Cassini spacecraft (with 12 instruments) with its companion, the Huygens probe (with 6 instruments), was launched from Cape Canaveral on 15 October 1997. The massive mission, about 5650 kg, needed three major gravity assists and 7 years to reach Saturn. It looped around the Sun twice, boosting its orbit for a final gravity assist from a close flyby of Jupiter in December 2000. During the Jupiter flyby, the Cassini instruments were turned on to measure properties of the jovian system, with spectacular results.



After the flyby, Cassini began its fall toward the ringed planet. On 11 June 2004, Cassini flew within 2071 km of the outermost moon of Saturn, Phoebe: its closest and only approach to this satellite. Phoebe is covered with craters signifying an ancient surface and consists of a mixture of rock and water ice with an unexpectedly high porosity. On 1 July 2004 Cassini fired its main rocket engines to slow down the spacecraft, so it could drop into a stable orbit. As the spacecraft was slowing down, it glided over the rings and obtained unprecedented details of their structure and chemistry during the only close ring plane crossing of the entire mission. New moons were found within ring gaps; new rings were mapped; fine-scale ring structures were puzzled over; clouds of oxygen suggested tenuous ring atmospheres; and other data suggested that the rings formed from recycled bits of the moons, making the rings second- or perhaps even third-generation offspring of their parent planet.

During the ring plane crossing, Cassini also made its closest approach to Saturn, coming within 18,000 km of the gas giant. Before, during, and after this close approach, instruments tracked lightning associated with storms, clouds, vertical wind shears, and thermal variations in the atmosphere. These observations enigmatically imply that Saturn's rotation rate has slowed by about 6 min since 1981. Saturn's atmosphere is richer in carbon than Jupiter's atmosphere, which is consistent with planet formation by core accretion. A radiation belt was discovered inside of the D ring, the external magnetic field has changed, the current sheet within the magnetosphere is thinner and more extended, and ion cyclotron waves indicate a dynamic plasma.

After Cassini's closest approach to Saturn, the spacecraft dropped down through the large gap between the F and the G rings into a stable orbit. The size and angle of this orbit will change over time to allow Cassini to visit many moons and sample the rest of the saturnian system to develop a more complete three-dimensional model of the system's dynamics. Cassini plans to do more than drop by; it will stay awhile.

–LINDA ROWAN

CONTENTS

Science

Published by AAAS

How Long Is the Day on Saturn?

Agustín Sánchez-Lavega

Determining a planet's rotation period can be difficult if the planet lacks a solid surface. However, for planets with an internal magnetic field, emissions at radio wavelengths are modulated by the planet's rotation rate. The latest results from the Cassini spacecraft seem to indicate that Saturn's rotation rate has slowed down by 6 minutes since the Voyager 1 and 2 spacecraft flew by the planet in 1980 and 1981, but it is unclear whether a slowdown has in fact occurred. Future data collected by Cassini may be able to resolve the question.

Observations of Saturn's radio emission by the Cassini spacecraft (*1*) confirm previous findings by the Ulysses spacecraft (*2*): The periodicity of the emission has changed by more than 6 min since Voyager 1 and 2 flew by the planet in 1980 and 1981 (*3*, *4*). The periodicity of the radio emission is assumed to represent the planet's rotation rate. A fundamental question thus arises: How long is the day on Saturn?

There are two ways to determine the rotation rate of a planet or other astronomical body without a solid surface. One way is to measure the motions of features on the visible part of the atmosphere; however, such motions may not reflect the rotation of the planet itself. The alternative is to track the rotational modulation of the radio emission. Not all radio emissions are linked to the magnetic field, but for those that are, the radio period is a measure of the rotation rate of the body because the magnetic field originates in the interior of the body. The approach has been particularly useful for the giant planets Jupiter, Saturn, Uranus, and Neptune, which lack an accessible surface and have atmospheric motions that strongly depend on latitude.

Saturn lacks a modulated radio signal that can be detected from Earth. Therefore, few data on the rotation period of Saturn were available before the Voyagers flew by the planet. From 1876 to 1978, ground-based telescopes were used to track the motions of a dozen or so cloud features, yielding an atmospheric differential rotation rate of 10 hours 14 min at the equator and 10 hours 42 min at other latitudes (*5*).

The situation changed when the Voyagers' antennas detected a radio emission that allowed the radio rotation period to be determined. The emission was in the frequency range from about 50 to 500 kHz and is called Saturn's kilometric radio (SKR) emission. Its temporal modulation yielded a rotation rate of 10 hours 39 min 22.4 s (±7 s) (Fig. 1); this is the value recommended by the International Astronomical Union (*6*). The emission was confined to small regions above Saturn's polar zones in both hemispheres (latitudes from ~60° to ~80°) and is apparently generated by instabilities on auroral field lines.

But this new rotation rate was soon questioned by planetary meteorologists in view of the large wind speeds at the equator (close to 500 m/s, or about two-thirds of the speed of sound in Saturn's light and cold atmosphere). These wind speeds are determined relative to the planet's rotation rate; very fast eastward jets (*7*) are difficult to reproduce in atmospheric circulation models. If the planet's rotation rate were lower, the atmospheric jet patterns would be more compatible with those seen on Jupiter and with model proposals for atmospheric circulation on giant planets.

Matters became even more complicated when Ulysses observations of SKR emissions between 1994 and 1997 showed changes of up to 1% in the rotation period relative to the Voyager observations, with a new average radio period of 10 hours 46 min (*2*). This value is confirmed by the more precise Cassini value of 10 hours 45 min 45 s (±36 s) (*1*). Given these disparate results, does the SKR emission really represent Saturn's rotation? If so, what caused the observed period drift between 1980–1981 and 1994–2004?

It is difficult to believe that Saturn's rotation period has really changed, because angular momentum conservation would require a large change in the interior angular velocity or in the inertial moment of the planet to explain the 6-min difference. One possibility is that Saturn's interior rotates not as a solid body but differentially, as occurs in the in-

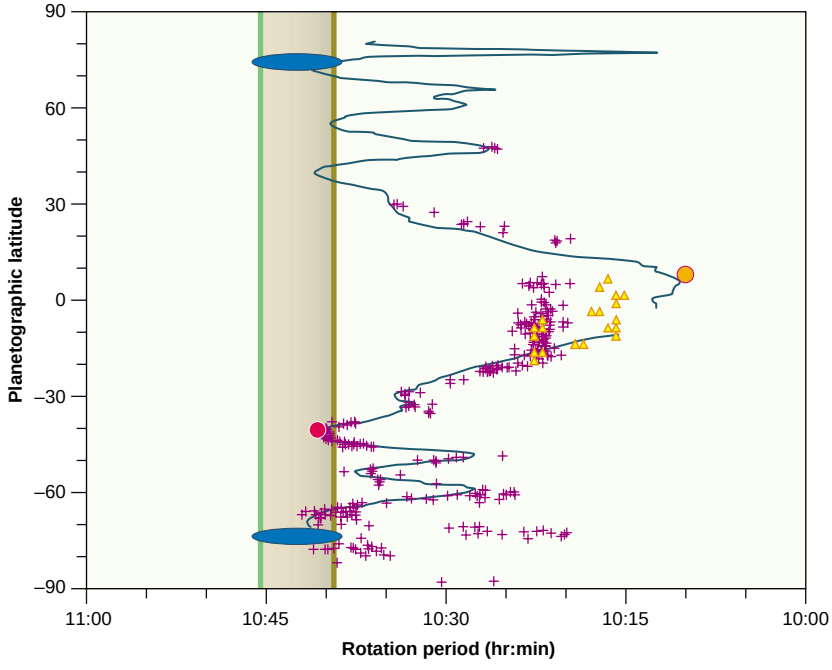


Fig. 1. Saturn's rotation periods. Cloud tracking of atmospheric features, as measured during the Voyager 1 and 2 encounters in 1980 and 1981, are represented by a dark continuous line (*15*). Violet crosses mark points measured from 1994 to 2004 with the Hubble Space Telescope (*9*). Yellow triangles are the equatorial data measured by Cassini (*10*). Outside the equator, the Voyager, Hubble, and Cassini data essentially coincide. The dots indicate short radio emission bursts that have been assigned to lightning produced by storms [yellow dot, Voyager (*16*); red dot, Cassini (*1*, *10*)]. The brown vertical line is the SKR emission period from Voyager (*3*, *4*, *6*), the green vertical line is the SKR emission as measured by Cassini (*1*), and the shaded area indicates the periods determined by the Ulysses spacecraft (*2*). The blue ovals indicate the approximate latitudes of the SKR emissions.

Escuela Superior de Ingeniería, Universidad del País Vasco, 48013 Bilbao, Spain. E-mail: wupsalaa@bi.ehu.es

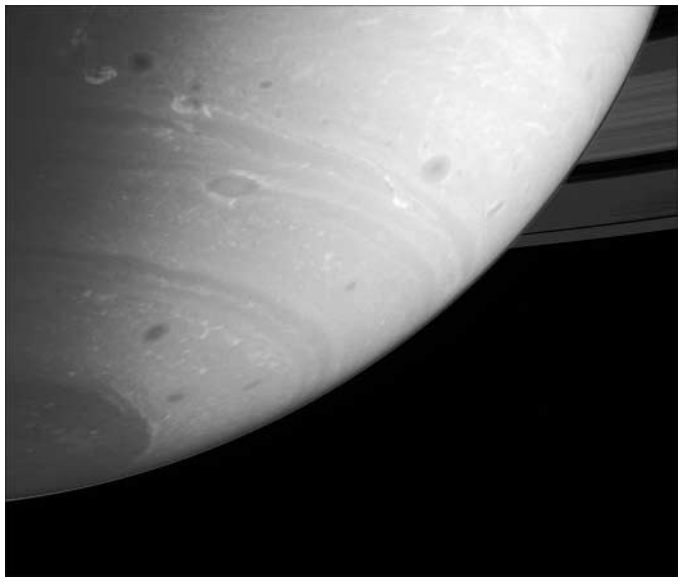


Fig. 2. Saturn's atmosphere. This Cassini image of the planet's southern hemisphere shows dark storms ringed by bright clouds. It was taken with the Cassini narrow-angle camera on 19 September 2004 at a distance of 8.3 million km through a filter sensitive to infrared light at 750 nm. [Image: NASA/ESA]

terior of the Sun (*8*), and that we detect radio emissions from different regions in the planet. But there is at least one other possibility.

The period change in the SKR emission has coincided with a large variability in the speed of the equatorial jet at cloud level (*9*). Two reports in this issue add new data on winds at different altitudes in Saturn's atmosphere, one tracking cloud motions (*10*) (Fig. 2) and the other temperatures (*11*). This wind variability might be due to a global circulation change, a vertical wind shear effect, the motion of atmospheric waves that do not transport matter, or a combination of these factors. Perhaps the strong storm activity displayed at cloud level in equatorial latitudes since 1990 (*12*) has influenced deeper atmospheric motions. The latter may in turn affect the magnetic field period if the two are coupled, as some models suggest (*13*).

However, nobody has previously predicted that huge storm activity could change the rotation rate of a planet's magnetic field; furthermore, the physics involved are far from understood. Most probably there will be a simpler explanation that relies on a better understanding of how the SKR emission is generated and how the magnetic field works on Saturn.

Saturn's rotation rate is a fundamental parameter for atmospheric, magnetospheric, and interior studies of the planet. For example, planetary meteorologists need a frame of reference for atmospheric wind speeds. Determination of a definitive, fixed rotation period is therefore a high priority. The solution could arrive from two sets of measurements.

First, the rotation rate of the magnetic field could be measured directly by the Cassini magnetometer while the spacecraft is in orbit. In fact, the Pioneer 11 and Voyager 1 magnetometers detected a periodicity in the azimuthal component of the magnetic field that was consistent with the SKR emission period at that time (*14*). Second, long-term measurements of the intensity, source location, periodicity, and fluctuations of SKR emissions could provide the answer. Cassini's instruments can fulfill both tasks as the spacecraft orbits the planet for the next 4 years.

References

1. D. A. Gurnett *et al.*, *Science* **307**, 1255 (2005); published online 16 December 2004 (10.1126/science.1105356).
2. P. H. M. Galopeau, A. Lecacheux, *J. Geophys. Res.* **105**, 13089 (2000).
3. M. D. Desh, M. L. Kaiser, *Geophys. Res. Lett.* **8**, 253 (1981).
4. T. D. Carr *et al.*, *Nature* **292**, 745 (1981).
5. A. Sánchez-Lavega, *Icarus* **49**, 1 (1982).
6. P. K. Seidelmann *et al.*, *Celest. Mech. Dyn. Astron.* **82**, 83 (2002).
7. M. Allison, P. H. Stone, *Icarus* **54**, 296 (1983).
8. A. J. Dressler, *Geophys. Res. Lett.* **12**, 299 (1985).
9. A. Sánchez-Lavega *et al.*, *Nature* **353**, 397 (1991).
10. C. C. Porco *et al.*, *Science* **307**, 1243 (2005).
11. F. M. Flasar *et al.*, *Science* **307**, 1247 (2005); published online 23 December 2004 (10.1126/science.1105806).
12. A. Sánchez-Lavega *et al.*, *Nature* **423**, 623 (2003).
13. R. L. Kirk, D. J. Stevenson, *Astrophys. J.* **316**, 836 (1987).
14. G. Giampieri, M. K. Dougherty, *Geophys. Res. Lett.* **31**, 16701 (2004).
15. A. Sánchez-Lavega, P. V. Sada, J. F. Rojas, *Icarus* **147**, 405 (2000).
16. M. L. Kaiser *et al.*, *Nature* **303**, 50 (1983).

10.1126/science.1104956

VIEWPOINT

Saturn's Variable Magnetosphere

Tamas I. Gombosi* and Kenneth C. Hansen*

Since the Cassini spacecraft reached Saturn's orbit in 2004, its instruments have been sending back a wealth of data on the planet's magnetosphere (the region dominated by the magnetic field of the planet). In this Viewpoint, we discuss some of these results, which are reported in a collection of reports in this issue. The magnetosphere is shown to be highly variable and influenced by the planet's rotation, sources of plasma within the planetary system, and the solar wind. New insights are also gained into the chemical composition of the magnetosphere, with surprising results. These early results from Cassini's first orbit around Saturn bode well for the future as the spacecraft continues to orbit the planet.

The late Fred Scarf, a pioneer of the U.S. space program, said that in exploring other planets, "half the fun is just being in orbit." After more than a decade of preparation and a journey of nearly 7 years, the fun can begin for Cassini: The spacecraft, which carries the most sophisticated instrument package ever flown to a planetary magnetosphere (*1*), has reached Saturn's orbit. Several papers in this issue report Cassini's first close look at Saturn's magnetosphere (*2*–*7*).

The magnetosphere is the region dominated by the magnetic field of a planet. This field is strongly affected by the solar wind, a supersonic magnetized plasma that flows outward from the Sun and fills the interplanetary space. When the solar wind passes a magnetized planet, it severely compresses the field on the dayside of the planet and draws it out into a long "magnetotail" on the nightside. The solar wind not only confines the planetary magnetic field; it also transfers mass, momentum, and energy to the planet's magnetosphere, ionosphere, and upper atmosphere. The surface separating the interplanetary plasma from the

Center for Space Environment Modeling, Department of Atmospheric, Oceanic and Space Sciences, University of Michigan, Ann Arbor, MI 48109, USA.

*To whom correspondence should be addressed. E-mail: tamas@umich.edu (T.I.G.), kenhan@umich.edu (K.C.H.)

confined planetary magnetic field is called the magnetopause. In front of this surface, a shock wave (the bow shock) forms to deflect the solar wind around the magnetopause.

The magnetosphere extends 20 to 25 Saturn radii toward the Sun, enclosing the planet's rings and many of its moons. Between the edge of the main rings (3 Saturn radii) and about 9 Saturn radii, there are four moons (Mimas, Enceladus, Tethys, and Rhea) that are collectively referred to as icy satellites. Titan, Saturn's largest moon, orbits at 20 Saturn radii.

The magnetospheres of Jupiter and Saturn are more complex than that of Earth, for two reasons. First, these planets rotate much faster: The centrifugal acceleration is about three orders of magnitude larger at the surface of Jupiter and Saturn than on Earth. Second, they have large sources of plasma inside their magnetospheres. In the jovian magnetosphere, the volcanic moon Io produces about 10^{28} heavy ions per second. In the saturnian system, the rings and icy satellites add a substantial (and probably comparable to the jovian system) amount of heavy ions to the magnetosphere, with smaller additional contributions from the planet itself and its largest moon, Titan (*2*, *8*). This mass addition process is referred to as mass loading.

The initial Cassini measurements reported in this issue (*2*–*7*) reveal a highly dynamic magnetosphere. Some features remain unchanged since the first measurements more than 25 years ago. Saturn's internal magnetic field, as measured by the onboard magnetometer during Cassini's closest approach to the planet, is confirmed to be very nearly axially symmetric (*3*). But the new data reveal some apparent differences in today's magnetosphere compared to the Voyager findings and have led to several exciting discoveries.

The first indication that Saturn and its magnetosphere might be different than observed by the two Voyager spacecraft came when Cassini was still several astronomical units from Saturn. (An astronomical unit is the distance from Earth to the Sun.) The RPWS plasma wave instrument on Cassini observed a modulation period of Saturn's kilometric radiation (electromagnetic radiation with a wavelength on the order of a kilometer) that was nearly 6.5 min longer than that observed by the Voyagers (*4*). The modulation period of Saturn's kilometric radiation is associated with the rotation of Saturn itself. This "slowdown" poses the first Cassini puzzle about Saturn's magnetosphere [see the Viewpoint by Sánchez-Lavega (*9*)].

Gurnett *et al*. (*4*) attribute the increased modulation period to the fact that Saturn's magnetic and rotation axes are so closely aligned that the wobble induced by the slight misalignment is too small to control the rotational modulation; second-order effects, such as a magnetosphere that rotates more slowly than Saturn's upper atmosphere (a process referred to as slippage), can then become important. In our opinion, the increased modulation period also implies that the total magnetospheric plasma source is larger today than it was during the Voyager encounters; the increased mass-loading rate increases the slippage of the magnetosphere with respect to the ionosphere. This view is consistent with data from Cassini's UVIS instrument (*2*).

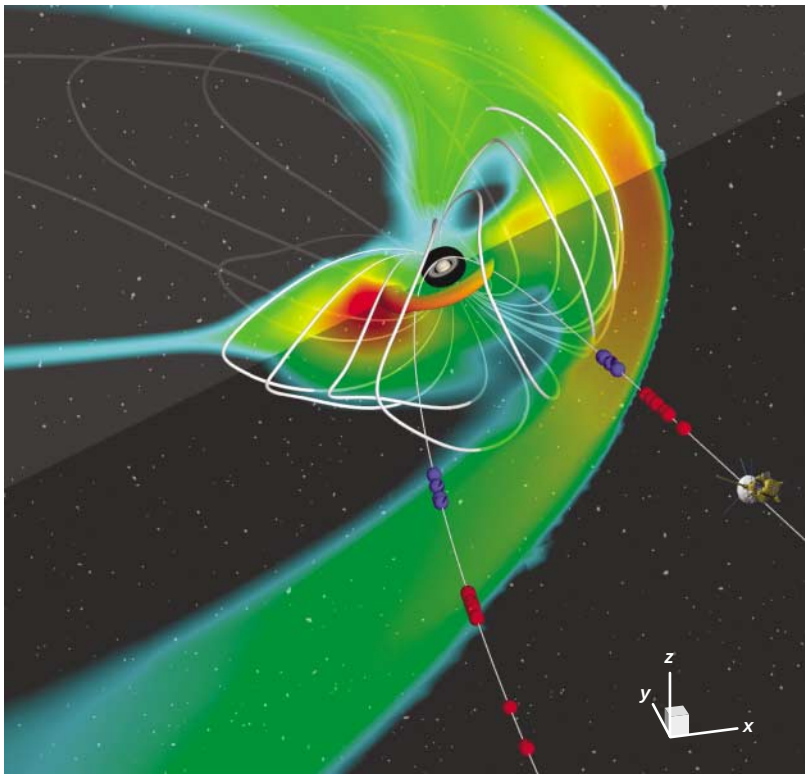


Fig. 1. A snapshot of Saturn's simulated magnetosphere. The solar wind is coming from the right. The thick white lines are a three-dimensional representation of the outermost magnetic field lines that have both ends connected to Saturn. On the dayside, they represent the magnetopause whereas on the nightside, they represent the magnetotail. The crescent at 6 to 7 Saturn radii shows that the plasma is concentrated near the rotational equator. The color code represents pressure in the equatorial and noon-midnight planes. Cassini's trajectory during its first, highly elliptical orbit is shown in gray, with observed bow shock (red) and magnetopause (blue) crossings. Multiple crossings occur because the bow shock and the magnetopause move in and out past Cassini as a result of changes in the dynamic pressure of the solar wind.

During its insertion into Saturn's orbit (Fig. 1), Cassini crossed the planetary bow shock some 17 times and encountered the magnetopause about 9 times (*3*–*5*). On the inbound leg of the orbit, the magnetosphere was more "inflated" than observed by the Voyagers (*3*). Data from the MAG instrument indicate that both the bow shock and magnetopause crossings occurred about 10 Saturn radii farther from Saturn than expected, possibly implying a much higher magnetospheric plasma content than was observed during the Voyager encounters.

In contrast, on the outbound leg the magnetopause was about 10 Saturn radii closer to the planet than anticipated. Two sets of bow shock crossings were observed on the outbound leg, one about 20 Saturn radii closer to the planet, the other near the expected location. The entire magnetosphere passage took about 400 hours (~17 days). The wide range of the bow shock and magnetopause crossings indicates that Saturn's magnetosphere is highly dynamic. This range can result from changes either in the internal plasma source or in the upstream solar wind conditions.

We recently carried out magnetohydrodynamic simulations of Saturn's magnetosphere (Fig. 1) (*10*). We used solar wind conditions corresponding to the observed plasma parameters during Cassini's orbital insertion and applied a magnetospheric plasma source that was close to the observed ~10^{28} ions/s (*2*). We found that the bow shock and magnetopause locations appear to be controlled by the dynamic pressure of the solar wind. In addition, the results indicate that Saturn's magnetosphere exhibits a quasi-periodic behavior. Neutral gas mainly originating from the rings and icy satellites is ionized in the inner magnetosphere and loads the closed magnetic field lines with an increasing mass of heavy ions. Centrifugal forces due to the rapid planetary rotation stretch the loaded magnetic field lines until their mass content reaches a critical value. Magnetic stresses cannot balance the centrifugal force, and the stretched field line pinches

and reconnects with itself somewhere closer to the planet on the dawn side of the magnetotail.

This process results in a short, closed magnetic field line that is nearly empty of plasma, and a closed magnetic loop that contains a large amount of heavy ions. The heavy, closed loop (the plasmoid) moves down the magnetotail and eventually leaves Saturn's magnetosphere. According to a preliminary analysis of the simulation results, the loading-unloading process is quasi-periodic, with plasmoids launching at multiples of Saturn's rotation period (the largest ejection occurs approximately every four rotations). This time period is probably controlled by the mass-loading rate and is modulated by the planetary rotation.

Cassini is the first planetary mission that can "image" a magnetosphere by detecting energetic neutral atoms created by charge transfer. It does so with the MIMI instrument, which has already made an exciting discovery: the existence of a radiation belt inside Saturn's innermost ring. This radiation belt is probably formed by the interaction between the main radiation belt and the upper layers of Saturn's exosphere. The nightside source of energetic neutral particles varies with an ~11-hour periodicity, whereas the source on the dayside remains approximately constant (*6*).

The vast improvement of Cassini's instruments over those on Voyager has resulted in a wealth of information about the plasma composition in the magnetosphere. Data from the high-resolution mass spectrometers CAPS and MIMI show that the inner magnetosphere is mainly dominated by hydrogen (H^+, H_2^+) and water-group ions (OH^+, H_2O^+, H_3O^+, O^+, O_2^+) (*5*–*7*).

Voyager's instruments could not differentiate water-group ions from nitrogen ions. With such differentiation now possible, the very low abundance of N^+ and N_2^+ in the magnetosphere is striking. Nitrogen ions escaping from Titan were expected to be the dominant species near Titan's orbit, but this is not the case. It remains unclear whether the low nitrogen abundance is due to reduced atmospheric escape from Titan or to a process that rapidly removes these ions from the magnetosphere.

Saturn's rings are one of the solar system's most spectacular phenomena. Cassini instruments have discovered a new element to this system: a tenuous atmosphere and ionosphere above the rings (*2*, *4*–*7*). The main constituent of the ring particles is assumed to be water ice, which produces a cloud of water products due to photo-processes and sputtering. Waite *et al*. suggest (*7*) that O_2 is the main constituent of the ring atmosphere; it can be produced by radiation-induced decomposition of ice. MIMI discovered a depletion of energetic ions and electrons inside Dione's orbit that is probably related to the higher-than-expected neutral gas density in this region. The most intriguing constituent of the ring-plane ionosphere is H_3O^+ (*5*). This ion is the major constituent of the dense ionospheres of comets and is produced via $H_2O^+ + H_2O \rightarrow H_3O^+ + OH$. The composition changes dramatically beyond ~9 Saturn radii to H^+-dominated plasma.

The new observations of Saturn's magnetosphere (*2*–*7*) paint a consistent picture of a highly dynamic magnetosphere dominated by inner plasma sources and planetary rotation but with substantial solar wind control. Cassini's first pass through the Saturn system has revealed a magnetosphere that, although familiar, holds many mysteries. The fun has begun.

References and Notes

1. M. Blanc *et al.*, *Space Sci. Rev.* **104**, 253 (2002).
2. L. W. Esposito *et al.*, *Science* **307**, 1251 (2005).
3. M. K. Dougherty *et al.*, *Science* **307**, 1266 (2005).
4. D. A. Gurnett *et al.*, *Science* **307**, 1255 (2005).
5. D. T. Young *et al.*, *Science* **307**, 1262 (2005).
6. S. M. Krimigis *et al.*, *Science* **307**, 1270 (2005).
7. J. H. Waite *et al.*, *Science* **307**, 1260 (2005).
8. S. Jurac, J. D. Richardson, *J. Geophys. Res.*, in press.
9. A. Sánchez-Lavega, *Science* **307**, 1223 (2005).
10. The simulations were carried out with the University of Michigan's three-dimensional magnetohydrodynamic code, BATS-R-US (*11*).
11. K. G. Powell *et al.*, *J. Comput. Phys.* **154**, 284 (1999).
12. This work was supported by the NASA Cassini Project under contract JPL-961178. The authors acknowledge the help of D. De Zeeuw in preparing the figure.

10.1126/science.1108226

RESEARCH ARTICLE

Cassini Imaging Science: Initial Results on Saturn's Rings and Small Satellites

C. C. Porco,[1]* E. Baker,[1] J. Barbara,[2] K. Beurle,[3] A. Brahic,[4] J. A. Burns,[5] S. Charnoz,[4] N. Cooper,[3] D. D. Dawson,[6] A. D. Del Genio,[2] T. Denk,[7] L. Dones,[8] U. Dyudina,[9] M. W. Evans,[3] B. Giese,[10] K. Grazier,[11] P. Helfenstein,[5] A. P. Ingersoll,[9] R. A. Jacobson,[11] T. V. Johnson,[11] A. McEwen,[6] C. D. Murray,[3] G. Neukum,[7] W. M. Owen,[11] J. Perry,[6] T. Roatsch,[10] J. Spitale,[1] S. Squyres,[5] P. Thomas,[5] M. Tiscareno,[5] E. Turtle,[6] A. R. Vasavada,[11] J. Veverka,[5] R. Wagner,[10] R. West[11]

Images acquired of Saturn's rings and small moons by the Cassini Imaging Science Subsystem (ISS) during the first 9 months of Cassini operations at Saturn have produced many new findings. These include new saturnian moons; refined orbits of new and previously known moons; narrow diffuse rings in the F-ring region and embedded in gaps within the main rings; exceptionally fine-scale ring structure in moderate– to high–optical depth regions; new estimates for the masses of ring-region moons, as well as ring particle properties in the Cassini division, derived from the analysis of linear density waves; ring particle albedos in select ring regions; and never-before-seen phenomena within the rings.

Major objectives of the Cassini Imaging Science investigation were the search for new satellites of Saturn within and external to the rings and determination of their orbits; refinement of the orbits of the known satellites; investigation of the temporal evolution of all satellite orbits; search for ring spokes; characterization of ring color and particle photometric properties; and the investigation of ring structure, dynamics, and temporal evolution, including interactions between rings and moons (*1*).

From 9 February 2004 to 1 June 2004, Cassini's narrow angle camera (NAC) (*1*) took ~800 images in four separate sequences designed specifically to search for new moons among the major icy saturnian satellites. A fifth sequence of 94 wide-angle frames designed to simultaneously image faint rings and search for new moons among the known moons was taken in October 2004 after Cassini entered orbit (*2*) (table S1).

Two new satellites orbiting between Mimas and Enceladus, S/2004 S1 (provisionally named Methone) and S/2004 S2 (provisionally named Pallene), were first seen in 1.2- and 4.6-s exposures taken on 1 June 2004 (*3*). [Contrary to earlier speculation (*3*),

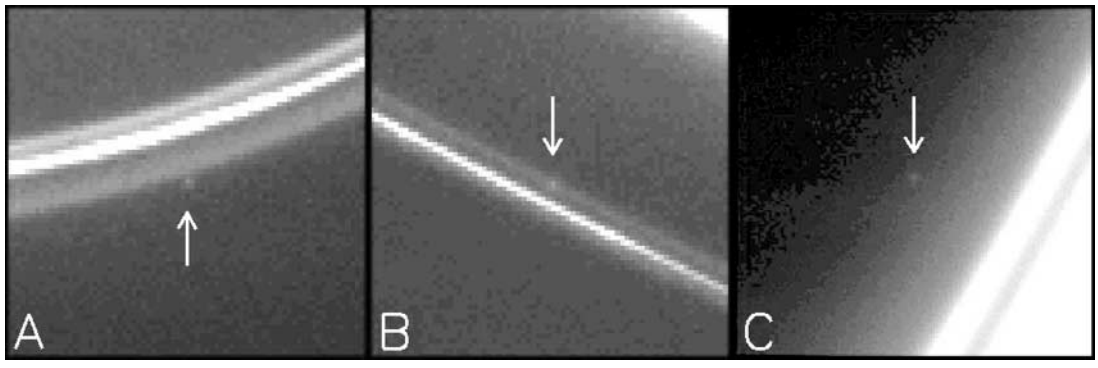


Fig. 1. Contrast-enhanced images showing newly discovered F-ring objects S/2004 S3 (**A**), S/2004 S4 (**B**), and S/2004 S6 (**C**). S/2004 S3 is shown in image N1466489981 taken on 21 June 2004 with a radial scale of 55 km/pixel and a phase angle of 66°. S/2004 S4 is shown in image N1466513582 taken on 21 June 2004 with a radial scale of 59 km/pixel and a phase angle of 66°. S/2004 S6 is shown in image N1477693877 taken on 28 October 2004 with a radial scale of 4.3 km/pixel and a phase angle of 151°.

it has been determined that the new moon Pallene is the same body as S/1981 S14, which was seen in one Voyager 2 image.] A third main-satellite region moon, S/2004 S5 (provisionally named Polydeuces), was found in a series of 5.6-s exposures in the fifth sequence (*4*). It is a trailing co-orbital of Dione.

Satellite searches were also conducted within the main ring system, including the F ring region. Three new objects were sighted. Two objects (S/2004 S3 and S/2004 S6) were detected in ring ansa movie sequences taken during June and October; another, S/2004 S4, was found in a ring photometry sequence, also on 21 June 2004, taken 5 hours after the first sighting of S/2004 S3 (*4*, *5*) (Fig. 1). The object S/2004 S3 clearly orbits outside the core of the F ring and outside the outer F-ring strand; S/2004 S4 and S/2004 S6 orbit interior to the F ring, the former within the inner F-ring strand and the latter interior to it. It is not clear at present whether these objects are moons or clumps (*6*).

No disk-resolved images of any of the new satellites have yet been taken. Size estimates have been made by assuming appropriate disk-integrated reflectivities (*7*) and taking into account the solar phase angle and the pixel size in kilometers to compute observed *I*/*F* values and subsequently derive a surface area (*2*). This calculation yielded approximate diameters of 3, 4, and 3.5 km for Methone, Pallene, and Polydeuces, respectively; S/2004 S3, S/2004 S4, and S/2004 S6 were found to be ~5 km across, assuming they are moons. Based on this method, Atlas, which orbits closely outside the A ring, and Pan, which orbits within the 325-km-wide Encke gap in the outer A ring, were found to be 20 and 25.6 km across, respectively, contrary to Voyager results (32 and 20 km, respectively) that suggested that Atlas was the larger of the two.

The objects S/2004 S3, S/2004 S4, and S/2004 S6 occupy a region near the F ring that is expected to produce chaotic orbits (*8*, *9*). No orbital integrations of satellites and ring particles in this region, aimed at examining the stability of these new bodies, have yet been performed.

The dynamical stability, and even the existence, of small objects like Methone, Pallene, and Polydeuces orbiting among the major saturnian satellites provide insight into their dynamical evolution and the history of impactors within the saturnian system. The result of a 4800-year integration of orbital motions of test particles in the vicinity of Mimas (*2*), (fig. S1), five times longer than previous integrations (*10*), shows that the region close to Mimas is very perturbed, and particles within ~0.2*R*s of Mimas are excited into eccentric and inclined orbits. (1*R*s is one Saturn radius, equal to 60,330 km.) Methone, very close to Mimas's region of influence, might be expected to have a substantial eccentricity and inclination, yet it does not (Table 1). Pallene is in a quiescent region and might be expected to have a smaller eccentricity and inclination, yet it does not.

A longer data arc than currently available will likely be necessary to reveal the effects of Mimas's perturbation on Methone. However, several explanations are possible for the eccentric and inclined orbit of Pallene: (i) a secular gravitational effect that is not manifest in only ~4800 years' integration; (ii) the moon migrated across eccentricity- and inclination-exciting resonances to its present position; or (iii) it is a fragment of a parent body that was hit by an impactor on an eccentric orbit. The third scenario could indicate possible heavy cometary bombardment in the past (*2*).

Many Cassini images were planned to refine the orbits of the known small moons; many other images targeted to Saturn or the rings capture these bodies serendipitously. Once their orbits were determined, new moons discovered in satellite searches were recovered in subsequent retargeted observations (*4*). Positional measurements of stars and moons made in these Cassini images plus an accurate description of the spacecraft trajectory form the basic data used to perform image navigation and determine the orbits of all satellites over the time interval spanned by the data.

The classical elements for the eight very small (<25-km diameter) satellites observed by Cassini ISS appear together with their standard errors in Table 1. The model for most of the satellite orbits is a Keplerian ellipse referred to the latest Cassini-era Saturn equator with apse and node rates determined by Cassini-era values for Saturn's gravitational harmonics (*2*). [Where Voyager data were available for Pan and Atlas, these have been used in the orbital fits (*2*).] However, because Polydeuces was found librating around Dione's trailing L5 Lagrangian point, its orbit was determined by performing a numerical integration to all the observations, including the perturbations due to the major saturnian satellites and the oblateness of Saturn. Polydeuces is librating ~68° behind Dione, with an amplitude of 25.8° and a period of 792 days. Dione's previously known Lagrangian satellite, Helene, leads Dione by ~62° and librates with an amplitude of 14.8° and a period of 769 days (*11*). These amplitudes are considerably larger than the 1.3° and 3.6° libration amplitudes of the leading and trailing co-orbitals of Tethys, Telesto, and Calypso. A set of mean orbital elements was developed for Polydeuces by fitting a precessing ellipse to the integration over the time span 1 January 2004 to 1 January 2009, but correcting the mean longitude to account for the libration.

[1]Cassini Imaging Central Laboratory for Operations, Space Science Institute, 4750 Walnut Street, Suite 205, Boulder, CO 80301, USA. [2]Goddard Institute for Space Studies, NASA, 2880 Broadway, New York, NY 10025, USA. [3]Astronomy Unit, Queen Mary, London E1 4NS, UK. [4]Centre d'Etudes de Saclay, Université Paris 7, L'Orme des Merisiers, 91191 Gif sur Yvette Cedex, France. [5]Department of Astronomy, Cornell University, Space Sciences Building, Ithaca, NY 14853, USA. [6]Department of Planetary Sciences, University of Arizona, 1629 East University Boulevard, Tucson, AZ 85721, USA. [7]Institut für Geologische Wissenschaften, Freie Universität, 12249 Berlin, Germany. [8]Department of Space Studies, Southwest Research Institute, 1050 Walnut Street, Suite 400, Boulder, CO 80302, USA. [9]Division of Geological and Planetary Sciences, California Institute of Technology, 150-21, Pasadena, CA 91125, USA. [10]Institute of Planetary Research, German Aerospace Center, Rutherfordstrasse 2, 12489 Berlin, Germany. [11]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA.

*To whom correspondence should be addressed. E-mail: carolyn@ciclops.org

Table 1. Planetocentric elements of the small saturnian moons. The epochs are given as Julian ephemeris dates. A sufficient number of observations of all satellites, with the exception of S/2004 S4, were available to permit the determination of a full set of elements for each. For S/2004 S4, we fit a circular equatorial orbit to the observations. Polydeuces is a trailing co-orbital of Dione with a large amplitude of libration. The set of mean orbital elements given here for Polydeuces was developed by fitting a precessing ellipse to the integration over the time span 1 January 2004 to 1 January 2009 and correcting the mean longitude to account for the libration. We compute the values of a, $d\varpi/dt$, or $d\Omega/dt$ from the formulae given in (*75*); their uncertainties stem from the errors in the satellites' mean motions and the error in the second zonal harmonic of Saturn's gravity field.

Element	Atlas (SXV)	Pan (SXVIII)	Methone (S/2004 S1)
Epoch	2,453,177.5	2,453,177.5	2,453,177.5
a (km)	137,665.0 ± 1.0	133,584.0 ± 1.0	194,251.0 ± 1.0
e ($\times 10^3$)	1.20 ± 0.06	0.21 ± 0.08	1.0 ± 0.2
i (°)	0.009 ± 0.007	0.007 ± 0.002	0.018 ± 0.007
λ (°)	11.398 ± 0.008	146.611 ± 0.008	194.76 ± 0.02
ϖ (°)	82.3 ± 3.6	249.0 ± 10.0	348.0 ± 18.0
Ω (°)	281.0 ± 27.0	303.0 ± 40.0	351.0 ± 29.0
$d\lambda/dt$ (°/day)	598.313997 ± 0.000003	626.031719 ± 0.000004	356.5365 ± 0.0002
P_λ (s)	51,986.0811 ± 0.0003	49,684.3834 ± 0.0003	87,239.31 ± 0.4
$d\varpi/dt$ (°/day)	2.8812 ± 0.0012	3.2068 ± 0.0014	0.8540 ± 0.0003
$d\Omega/dt$ (°/day)	−2.8674 ± 0.0012	−3.1905 ± 0.0014	−0.8519 ± 0.0003
Element	Pallene (S/2004 S2)	S/2004 S3	S/2004 S4
Epoch	2,453,177.5	2,453,177.5	2,453,177.5
a (km)	212,283.0 ± 1.0	141,071.0 ± 1.0	140,171.0 ± 41.0
e ($\times 10^3$)	4.0 ± 0.2	4.0 ± 0.2	0.0
i (°)	0.181 ± 0.008	0.05 ± 0.02	0.0
λ (°)	125.48 ± 0.05	29.90 ± 0.02	27.9 ± 0.1
ϖ (°)	78. ± 7.0	154. ± 4.0	
Ω (°)	7.0 ± 6.0	188.0 ± 14.0	
$d\lambda/dt$ (°/day)	312.0271 ± 0.0009	576.7104 ± 0.0006	582.3 ± 0.3
P_λ (s)	99,683.6 ± 0.3	53,933.48 ± 0.05	53,417.0 ± 24.0
$d\varpi/dt$ (°/day)	0.6243 ± 0.0002	2.6414 ± 0.0008	
$d\Omega/dt$ (°/day)	−0.6230 ± 0.0002	−2.6295 ± 0.0008	
Element	Polydeuces (S/2004 S5)	S/2004 S6	
Epoch	2,453,006.5	2,453,177.0	
a (km)	377,390.0	140,760.0 ± 12.0	
e ($\times 10^3$)	18.2	7.5 ± 1.5	
i (°)	0.1705	0.02 ± 0.01	
λ (°)	107.58	27.0 ± 10.0	
ϖ (°)	143.19	240.0 ± 20.0	
Ω (°)	302	25.0 ± 5.0	
$d\lambda/dt$ (°/day)	131.53	578.63 ± 0.08	
P_λ (s)	236,468	53,755.0 ± 7.0	
$d\varpi/dt$ (°/day)	0.0889	2.6626 ± 0.0009	
$d\Omega/dt$ (°/day)	−0.0773	−2.6504 ± 0.0009	

Table 2. Planetocentric mean elements of Janus, Epimetheus, Prometheus, and Pandora at Julian ephemeris date 2453005.50 referred to the Saturn equator. These are the elements of a precessing ellipse fit to the integration from 1 January 2004 to 1 January 2005. This interval avoids the time of closest approach for Janus and Epimetheus (February 2006) and the time of anti-apse alignment for Prometheus and Pandora (late 2006). Thus, the quoted orbital elements are free of the effects of the gravitational interactions between the satellites. The mean orbits during this interval, however, exhibit periodic differences from the integrated ones, with the largest differences being in the along-orbit direction. The amplitudes of the differences are about 200, 900, 200, and 1000 km for Janus, Epimetheus, Prometheus, and Pandora, respectively. The integrated orbits themselves are presumed to be accurate to better than 150 km for Janus, Epimetheus, and Pandora and 300 km for Prometheus.

Element	Janus	Epimetheus	Prometheus	Pandora
a (km)	151,460.0	151,410.0	139,380.0	141,710.0
e	0.0067	0.0098	0.0023	0.0042
i (°)	0.1649	0.3541	0.0056	0.0522
λ (°)	171.4792	346.0186	306.1799	252.6616
ϖ (°)	288.2511	38.5429	63.2935	50.4554
Ω (°)	48.1413	85.5628	266.0754	329.9818
$d\lambda/dt$ (°/day)	518.238639	518.483719	587.284953	572.792057
P_λ (days)	0.6947	0.6943	0.6130	0.6285
$d\varpi/dt$ (°/day)	2.0528	2.0552	2.7573	2.5994
$d\Omega/dt$ (°/day)	−2.0447	−2.0472	−2.7445	−2.5877

The addition of the Cassini data has improved knowledge of the mean longitude rates of Atlas and Pan by four and three orders of magnitude, respectively, over their previously published values. No significant inclination has been detected in the orbit of Atlas. However, the orbits of both, especially that of Atlas, exhibit a measurable eccentricity and Pan looks to have a significant, though very small, inclination. There are no important resonances within a resonance width of Pan's orbit that might affect its orbit, and no other comparably sized moons have been found in the Encke gap. Hence, we conclude that Pan's eccentricity is derived from its interactions with the surrounding ring material, a result that is sure to shed light on the mechanism by which exosolar planets embedded in disks derive their eccentricities (*12*).

Table 2 contains a set of mean elements for the "co-orbital" moons, Janus and Epimetheus, and the F-ring shepherds, Prometheus and Pandora, derived from both Cassini and earlier measurements (*2*). Because of strong dynamical interactions between the members of each pair, the orbits of these moons were numerically integrated (*13*).

Saturn's rings were imaged regularly on approach to Saturn and then immediately following Saturn orbit insertion (SOI) when the spacecraft came within 16,000 km above the unilluminated side of the rings. The SOI darkside sequence resulted in image scales in the ring plane between ~100 and 350 m/pixel. This series of narrow-angle SOI-darkside images (which, because of spacecraft motion, does not form a complete, uninterrupted scan of the rings) begins in the outer C ring, crosses into the inner B ring, skips the middle B ring, and captures the outer B ring, Cassini division, and the breadth of the A ring. Immediately following this period, the spacecraft crossed the ring plane and acquired a series of dayside NAC and (simultaneous) WAC (wide angle camera) images, extending from the F ring inward and terminating just interior to the Encke gap. These SOI-dayside NAC images have radial scales of ~1 km/pixel, several times coarser than the dark-side images but with five times better resolution than anything acquired by Voyager. The total number of images acquired during the dark- and lit-side post-SOI period was 85.

Nine days before it entered orbit around Saturn, Cassini imaged the lit face of the main rings with the NAC in five broadband filters (*1*): UV3 (338 nm), BL1 (451 nm), GRN (568 nm), RED (650 nm), and IR2 (862 nm), at a phase angle of 66° and an image scale of 38 km/pixel (Fig. 2). We constructed ring spectra (Fig. 3), using overall, or Bond, albedos of the ring particles as a function of distance from Saturn and wavelength that we inferred by using a classical multiple-scattering code (*2*) (fig. S2).

Spectra of the rings in 12 regions similar to those considered by (*14*) show that the

largest variations in spectral slope occur at short wavelengths (*15*). The B ring is generally "redder" than the A ring, which, in turn, is redder than the C ring and Cassini division (Fig. 3) (*16*). Because the rings' scattering function is insensitive to the physical thickness of the rings in this intermediate-phase viewing geometry, our spectra provide evidence that most small-scale variations in ring brightness result from compositional differences (*17*, *18*). If the rings began as nearly pure water ice, regions of smaller optical depth are expected to become more polluted by interplanetary dust, resulting in reduced particle albedos (*19*, *20*).

Spokes, the transient, quasi-radial ghostly features observed by Voyager in Saturn's B ring (*21*) and seen also in Hubble Space Telescope (HST) images (*22*), were not seen in any Cassini image taken on approach to Saturn, during which the solar incidence angle was ~65° and the emission angle was ~73°. Radiative transfer modeling of HST observations of spokes indicated that they might be seen when either the spacecraft or Sun's elevation angle above the ring plane got below ~7° (*22*). Such a grazing viewing geometry occurred for Cassini on its first periapse passage by Saturn after SOI during October 2004. However, no spokes were observed at that time either. The creation and/or visibility of spokes may depend critically on the Sun's elevation above the rings. High Sun creates a layer of photoelectrons above the rings that can negatively charge small dust particles above the rings, pulling them back to the (positively) charged rings. A low solar elevation angle reduces the number density in the photoelectron layer, causing dust particles to have a net zero (or slightly positive) charge and therefore to be repelled by the positively charged ring (*23*). The relatively high Sun elevation at present may create an environment hostile to the appearance of spokes. If this is true, spokes may not become visible until later in the orbital tour when the Sun's elevation is lower.

The highest resolution images taken shortly after SOI have revealed new ring phenomena as well as new details at higher signal-to-noise (S/N) ratios than previously achieved on ring features—e.g., gap edges, narrow ringlets, spiral density and bending waves—first seen in Voyager occultation scans or images. Moreover, with two-dimensional images, one can confidently distinguish real, longitudinally extended ring features from noise, a problem that hampers the interpretation of low S/N Voyager occultation data taken in high–optical depth regions of Saturn's rings. Many of our findings now provide new insights into the ring particles' collective effects (e.g., collisions, viscosity, self-gravity) and how these effects are altered in the presence of gravitational perturbations by moons.

Two ring regions of interest, captured in the dark-side SOI images (*24*), are shown in

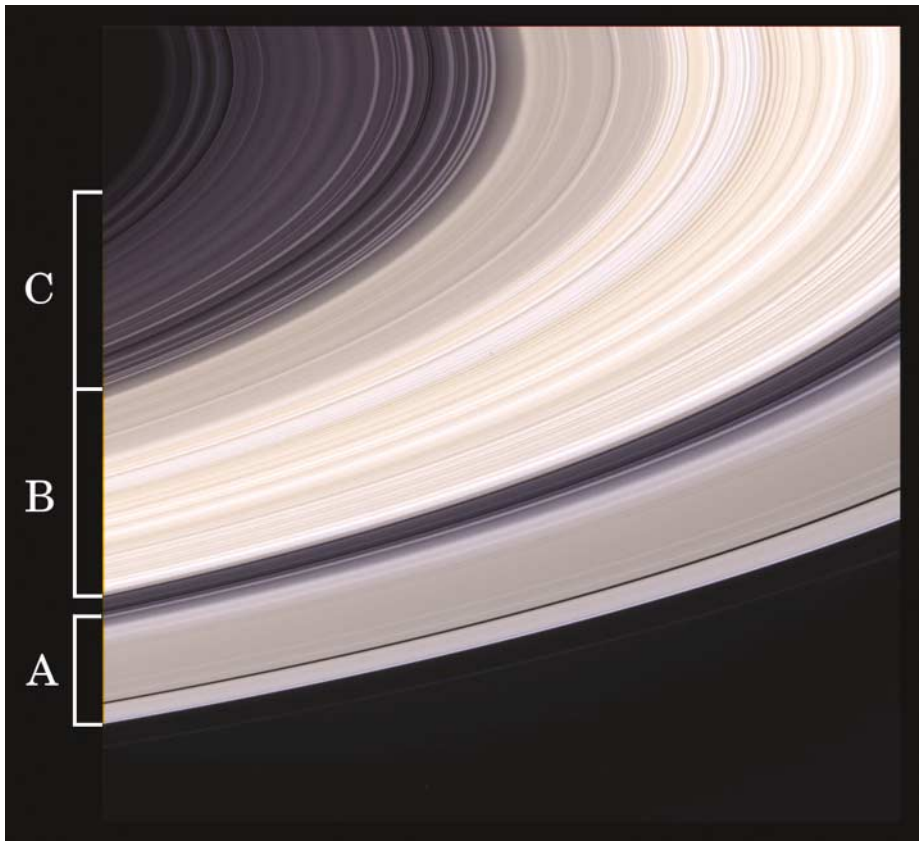


Fig. 2. Nine days before it entered orbit, Cassini captured this natural color view of Saturn's rings using the BL1 (451 nm), GRN (568 nm), and RED (650 nm) filters. The images that comprise this composition were obtained with the narrow-angle camera on 21 June 2004, from a distance of 6.4 million km from Saturn and a phase angle of 66°. The image scale is 38 km/pixel. Many bands throughout the B ring have a pronounced sandy color; other subtle variations in color are obvious. Color variations in Saturn's rings have previously been seen in Voyager and Hubble Space Telescope (HST) images. Cassini images show that color variations across the rings are more pronounced in this viewing geometry (i.e., at intermediate phase angle) than they were when seen from HST at phase angles no greater than 6°. The different colors in the rings reflect different amounts of contamination of the predominant constituent—water ice—by other materials such as silicates or organics.

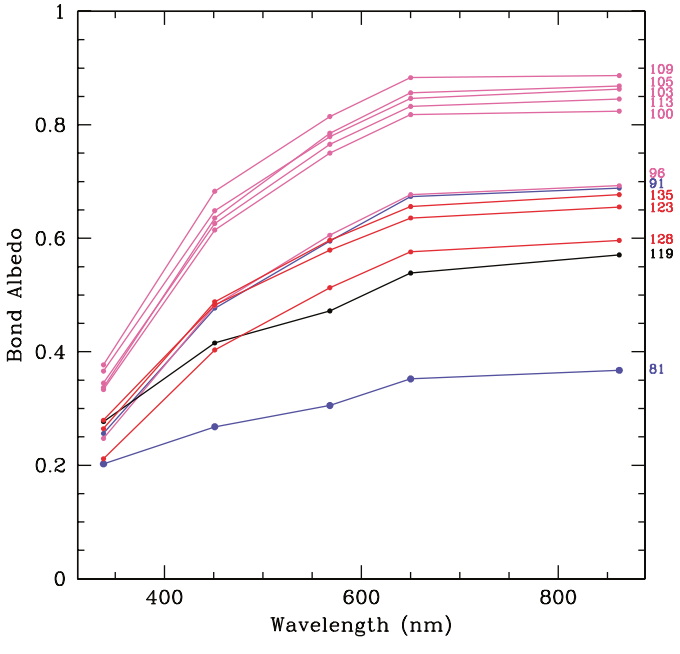


Fig. 3. Spectra of selected regions in Saturn's main rings. The main ring divisions are color-coded. This plot includes two regions in the C ring (blue), six regions in the B ring (green), one region in the Cassini division (black), and three regions in the A ring (red). Curves are labeled with the region's mean distance from Saturn's center in units of 1000 km.

Fig. 4. Figure 4A captures the narrow, eccentric Maxwell ringlet (*25*) in the gap of the same name at a resolution comparable to that of the Voyager stellar and radio occultation scans but at a higher S/N ratio. The Maxwell ring is analogous to the ε ring of Uranus in many of its physical and kinematic properties (*26*). A wave in the middle of the Maxwell ring, not discernible in the noisy Voyager occultation scans, is visible in the image and in a longitudinally averaged radial scan across the image (fig. S3a). A wave, presumably generated by the Cordelia 47:49 outer eccentric resonance (OER) (*27*), is observed in the middle of the uranian ε ring (*28*), but the wave in the Maxwell ring is not associated with any known resonance. No moon has yet been discovered in this gap either. The Maxwell ring wave is remarkably similar to waves produced by pulsation instabilities (also known as viscous overstabilities) in simulations of narrowly confined viscous rings (*29*), in which the narrow ring acts like a resonant cavity for amplifying a free wave disturbance. In this case, the pulsation frequency is comparable to the precession frequency and nonlocal transport of angular momentum dominates the viscous stresses. This suggestion can be verified by examining the correlation of vertical thickness of the ring with optical thickness in the wave in upcoming Cassini stellar and radio occultation data.

Figure 4B shows the narrow, eccentric Huygens ringlet on the left in the gap of the same name, a very narrow ringlet to the right of it, and one of the broad, low–optical depth ($\tau \sim 0.1$) plateaus for which the Cassini division is renowned. Figure S3b shows a longitudinally averaged radial scan of the Huygens ringlet seen in this image. This ringlet is known to have a very nonlinear width-radius relation (*30*). Here the ringlet is seen near its most narrow profile, and yet it appears near its most extreme radial extent. Cassini images reveal it to have a radial profile quite distinct from the Maxwell ringlet (fig. S3a), with optically thick ring edges. These observations may be related to the fact that the Huygens ringlet is the one narrow saturnian ring discovered so far to support two normal modes, one of which may be excited by a combination of the nearby Mimas 2:1 inner Lindblad resonance (ILR) and the eccentric outer B-ring edge (*30*, *31*). The optically thicker ring edges may also be the physical manifestation of the ring's self-gravity plus pressure forces enforcing rigid precession across the ring (*32*).

Comparing the physical characteristics and structures observed in different regions across the rings can help in disentangling the various contributing effects that generate ring structure and in determining the magnitude of their contributions. Two contrasting regions within the inner B ring, seen at radial scales of ~240 m/pixel, indicate the apparent sensitivity of ring-structure formation to the background optical depth. One region at 92,330 km (Fig. 5A) exhibits fine-scale irregular structure in a relatively high–optical depth ($\tau \sim 1.5$) part of the B ring, resembling in character the irregular structure seen in Voyager occultation and imaging data in equivalently high-τ parts of the B and A rings (*33*–*35*). However, irregular structure has never before been imaged at such a fine scale. Pulsation instability in dense rings may explain irregular structure on these length scales of 0.1 to 1 km (*36*). The second region in the B ring at 94,360 km (Fig. 5B), only ~2000 km away, with $\tau \sim 0.9$, is puzzling in its lack of any fine-scale structure and instead supports smooth undulations in brightness of ~100-km wavelength, with the bright bands in the image about 60 km across. These undulations are another form of unexplained structure and may be the result of a different physical mechanism. Ring regions of 100-km radial extent and reasonably high optical depth may exhibit sub-Keplerian shear rates because of the combined effects of finite particle size and physical adhesion between frost-covered particles. These regions would mimic an optically thick solid ring (*34*). Also, ballistic transport of material across ring regions resulting from meteoroid impacts (*33*) might also produce undulations of 100-km radial widths.

The outer portion of the B ring is among the densest in the rings ($\tau > 2$), and the $m = 2$ radial distortion in its outer edge is maintained by the strongest resonance—the Mimas 2:1 ILR—in Saturn's rings (*37*). In Cassini SOI images, the region 300 km interior to the outer edge at 117,273 km (Fig. 5C) also exhibits irregular structure at a scale not seen before. Such structure, as well as the orientation and amplitude of the $m = 2$ distortion (the latter is a factor of 3.5 larger than that predicted by test particle resonant dynamics), may be the result of wave reflection and pulsation instabilities at the outer B-ring edge (*38*).

The Cassini division, the region between the A and B rings that is comparable to the C ring in color and optical depth (Fig. 2), was imaged during the SOI sequence. One of its broad, relatively featureless plateaus, with a mean optical depth $\tau \sim 0.075$ (*39*), sports a very weak linear density wave (*40*), resolved here for the first time, created by the Atlas 5:4 resonance (Fig. 5D). The most optically thick region of the Cassini division (Fig. 5E) shows a wave feature caused by the Prometheus 5:4 ILR—a dark 180-km wide band—and two diffuse bright bands, each about 65 to 70 km across, reminiscent of the inner B-ring structure (Fig. 5C) with optical depths ~ 0.26. One of these bands also contains a weak Pan 7:6 ILR wave, which is not easily discernible in the image.

Analyses of the Pan and Atlas waves in the Cassini division have allowed determination of basic ring properties that had previously been determined with reasonable confidence only for the A and C rings. From a fit of the decrease of wavelength of the Atlas 5:4 wave with distance from resonance (fig. S4a), we infer a surface mass density $\sigma = 1.4 \pm 0.1$ g/cm^2. The quantity $R_{\mathrm{eff}} \equiv \sigma/(\tau\rho)$, where ρ is the internal density of a ring particle, is a measure of the effec-

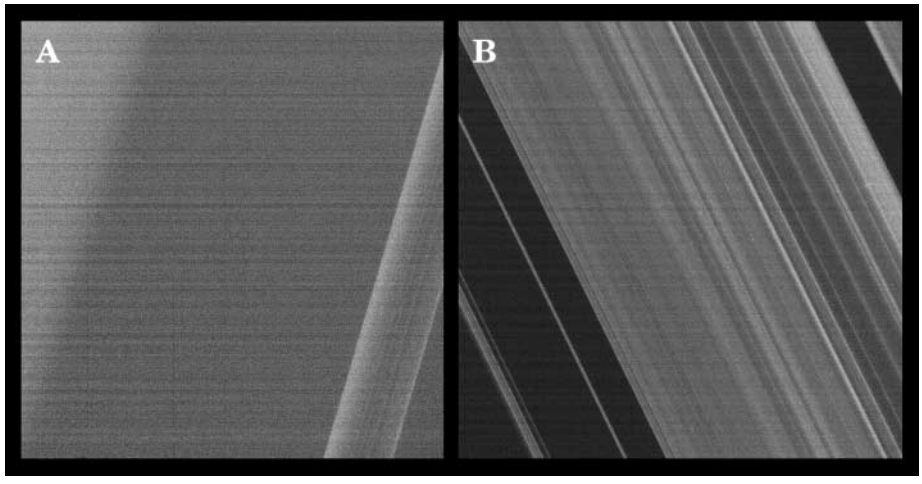


Fig. 4. Two SOI dark-side images of the Maxwell and Huygens ringlets, taken at a solar incidence angle of 114.5°. (**A**) The Maxwell ringlet is the band on the right, sitting in a gap of the same name. The image, N1467344214, is centered on 87,396 km, has a radial scale of 0.25 km/pixel, and was taken from a phase angle of 82° and an emission angle of 47°. A wave is faintly visible in the middle of it. A diffuse ringlet situated to the left of the Maxwell ringlet and discovered in later Cassini images (Fig. 7A) is not visible here. (**B**) The Huygens ringlet is the left-most narrow feature and appears alongside a very narrow ringlet and a broad plateau. This image, N1467345149, is centered at 118,009 km with a radial scale of 0.37 km/pixel and was taken from an emission angle of 63° and a phase angle of 59°. A diffuse narrow ringlet has also been found inward (to the left) of the Huygens ringlet (Fig. 7B), but is not visible here. Brightnesses measured across these ringlets have been converted to optical depth, τ, and compared with the Voyager PPS and RSS scans (*2*) (fig. S3).

tive size of the ring particles. Assuming ρ = 0.6 g/cm^3, we find $R_{eff} \sim 30$ cm at the location of the Atlas 5:4 wave, comparable to the value measured in the optically thin inner C ring (*41*) but nearly an order of magnitude smaller than we and others have found in Saturn's A ring (*42*). Thus the effective particle size is ~10 times larger in the A ring (*43*, *44*) than in the inner C ring and inner Cassini division. We also determined the viscosity, ν, for the Atlas 5:4 wave and found $\nu \sim 2$ cm^2/s. Using the "classical" expression $\nu = c^2\tau/[2\Omega(1 + \tau^2)]$ (*45*), where c is the dispersion, or random, velocity of the ring particles and Ω is their orbital frequency, this implies $c \sim 0.1$ cm/s and a ring scale height $H = c/\Omega \sim 6$ m.

We derived the first estimate of the mass of Atlas from the amplitude of the 5:4 wave; the "nominal" gravitational mass of Atlas is GM_{Atlas} = 0.00072 km^3/s^2 for an assumed density of 0.63 g/cm^3 and a mean diameter of 32 km based on Voyager images (*46*). However, our best fit gives GM_{Atlas} = 0.00014 km^3/s^2, i.e., a mass only ~20% of the nominal mass. As we discuss above, our modeling of unresolved Cassini images yields a diameter of 20 km for Atlas, implying a mean density for Atlas of 0.5 g/cm^3. This density agrees well with the densities of 0.4 to 0.6 g/cm^3 previously inferred for Epimetheus, Janus, Prometheus, and Pandora (*13*, *46*–*48*).

We observed a large number of density waves due to Pan in the SOI image sequence. Resolved waves range from the 7:6 in the Cassini division to the 90:89 in the mid-A ring. For the 7:6 wave in the outer Cassini division, we obtained a surface density σ = 2.5 ± 0.3 g/cm^2 from wave fits (fig. S4b). In this region, τ = 0.26; again assuming ρ = 0.6 g/cm^3 for the ring particles, we obtain $R_{eff} \sim 20$ cm, similar to the value we found in the inner Cassini division using the Atlas wave. The viscosity for the Pan 7:6 wave is 3.5 cm^2/s, implying $H \sim 5$ m.

Because of the larger optical depth of this region, determining the mass of Pan is not as straightforward as it is for Atlas. Even though the Pan wave is linear, the relation between the observed I/F variations and the optical depth variations in the wave is nonlinear. Thus, the optical depth (or surface mass density) profile we infer is model dependent. Our best fit gave GM_{Pan} = 0.00011 km^3/s^2, which corresponds to a density for Pan of only 0.19 g/cm^3 if we take Pan's mean diameter to be the Cassini value of D = 25.6 km, or a density of 0.39 g/cm^3 if we use Pan's pre-Cassini diameter of 20 km. Modeling of the Encke gap edge waves (see below), which is probably more reliable, gives a larger mass of GM_{Pan} = (0.00033 ± 0.00005) km^3/s^2, leading to a density of 0.56 g/cm^3 for D = 25.6 km or 1.18 g/cm^3 for D = 20 km. Most likely, Pan has a density smaller than that of solid ice (0.92 g/cm^2), like all the other satellites interior to Mimas with known densities.

The most optically thick region in the A ring (Fig. 5F), at a distance of 122,900 km and $\tau \sim 1.5$, shows irregular structure similar to that seen in other moderate– to high–optical depth regions in the rings but finer in scale. This is the finest structure seen anywhere in our images and may be evidence of the very long wave trains of undamped density waves, such as the Pandora 5:4 wave, which resides off the left-hand side of the image. A theoretical model of the damping of nonlinear density waves shows an abrupt transition in behavior with optical depth (*49*). Density waves in the moderate-τ outer A ring, such as the Mimas 5:3 wave, are predicted to have a wavelength (λ) that declines as x^{-1}, where x is the distance from resonance, as is observed, and also damp at roughly the predicted rate. Waves in slightly higher-τ regions, such as the inner B and A rings, are predicted to have $\lambda \sim x^{-1/3}$, contrary to observations, and to propagate essentially undamped forever. Although waves in high-τ regions, such as the Pandora 5:4, do propagate farther than waves in regions of smaller optical depth, all existing models underestimate damping of waves in high-τ regions (*49*–*51*).

The workings of strong satellite perturbations can be most clearly seen in SOI images of the A ring. A striking, uniformly spaced "corduroy" pattern overlain by two density waves is obvious in Cassini images (Fig. 5G) and provides the visual two-dimensional evidence for wakes excited by the satellite Pan (*52*, *53*) embedded in the Encke gap (just off

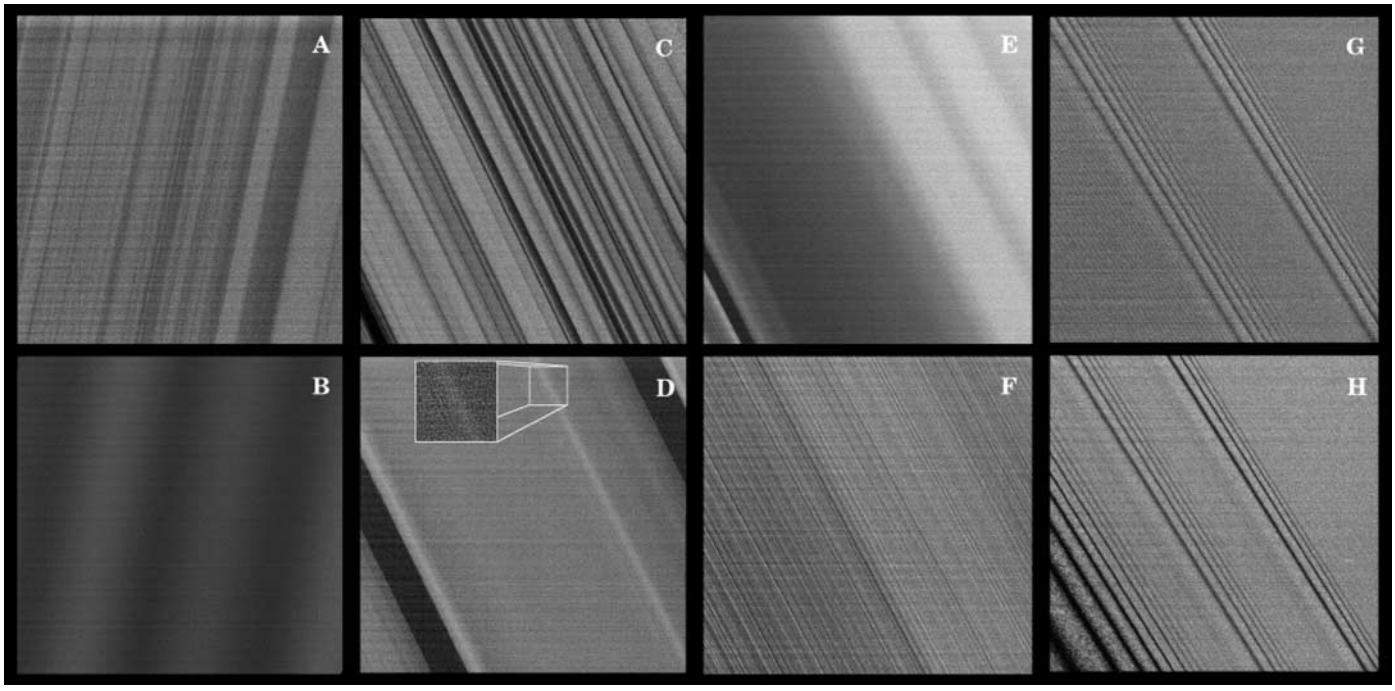


Fig. 5. This set of high-resolution images of Saturn's rings was taken during the dark-side SOI observation period with a solar incidence angle of 114.5°. Image scales range from 0.24 to 0.37 km/pixel in the ring plane. Images (**A**) and (**B**) were taken from an emission angle of 47° and a phase angle of 82°; images (**C**) through (**H**) were taken from an emission angle of 63° and a phase angle of 59°. Image names and radial locations of the centers of the images are as follows: (A) N1467344509, 92,326 km; (B) N1467344627, 94,362 km; (C) N1467345090, 117,273 km; (D) N1467345208, 118,745 km; (E) N1467345326, 120,496 km; (F) N1467345503, 122,893 km; (G) N1467346329, 133,227 km; and (H) N1467346447, 134,531 km.

the upper-right corner) less than ~200 km away. Wakes are created by the intersection of high–wave number, disturbed ring particle streamlines caused by the passage of Pan. The wakes in the "corduroy" pattern are produced in a manner identical to that of the waves seen on the inner edge of the Encke gap (see below), but here their wavelengths, amplitudes, and cant angles result from an encounter with Pan that was both more distant and further in the past. In fact, this region contains two wakes (fig. S5). The weaker, second-order wake was excited by Pan nearly two synodic periods earlier (1 synodic period = 0.39 years), but persists superimposed upon the first-order wake. Pan wakes cross the entire region and interact with a pair of density waves excited by moons exterior to the rings. The persistence of the wakes far from Pan and the inferred lack of damping may signify the importance of collective effects and self-gravity in maintaining coherent structure far from the moon.

A collection of new ring phenomena was seen in the dark-side SOI sequence that may be evidence of different physical manifestations of particle aggregation, caused by either gravitational instabilities or kinematical effects or both. The features known as "straw"—i.e., short-scale (a few kilometers), narrow, quasi-parallel variations in brightness—are seen in the first few troughs of strong density waves in dark-side SOI images of the rings. The most prominent examples are seen in the first troughs of the Janus 6:5 (Fig. 5H) and the Janus 4:3 density waves, as well as in the smaller Prometheus 33:32 wave ~55 km outside the Keeler gap. Other features, which show less prominent straw, are the Pandora 5:4, the Janus 2:1, the Pandora 9:8, the Mimas 5:3, and in the 3rd band/wake due to Pan ~13.5 km outside of the Encke gap. In most cases, straw is seen in several successive troughs of one wave, covering radial extents of dozens to greater than 100 km, fading out as the wave damps. One notable exception is the Prometheus 33:32 in which straw appears only in the first trough between the first two crests of the density wave. In all cases where straw is obvious, its longitudinal extent covers the entire field of view, which for dark-side images is less than 1°.

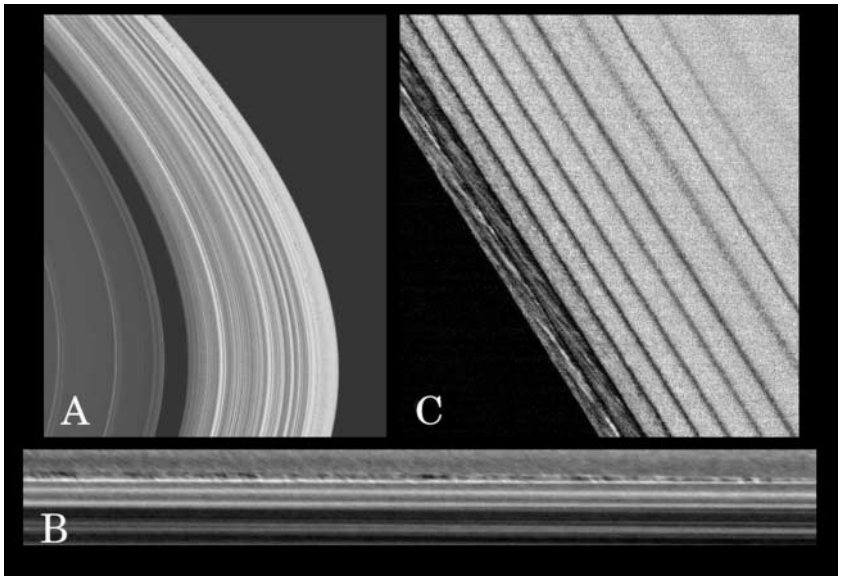


Fig. 6. Examples of two new ring phenomena: "mottled" and "ropy" structures. (**A**) Dayside SOI image of the outer edge of the A ring with solar incidence angle of 114.5°, an emission angle of 93.5°, and a phase angle of 129°. This image radial scale is 0.98 km/pixel. The longitudinal coverage is ~5.2°. (**B**) A map projection of the image (A), covering about 2.05° of longitude, roughly half of the full longitudinal extent seen in (A) and only 72 km in radial extent. (**C**) Dark-side SOI image N1467346388, with an incidence angle of 114.5°, an emission angle of 63°, and a phase angle of 59°, shows the outer edge of the Encke gap and the region exterior to it. The wakes of Pan are clearly seen. A different example of mottled structure is seen in the eighth Pan wake from the edge, as well as ropy structure within the first two bands exterior to the gap. This figure covers 180 km in radius and roughly 0.07° in longitude. The radial scale is 0.27 km/pixel and the longitudinal scale is 0.00013°/pixel.

Table 3. Measured radii and maximum brightnesses (*I/F*) for the diffuse ringlets seen in Fig. 7. "I" denotes inner; "O" denotes outer. The minimum optical depth needed to produce the *I/F* (assuming all forward-scattering dust), and the maximum optical depth needed to produce the *I/F* (assuming all large particles), are also given. These calculations assume that the observed *I/F* is due to single scattering. For R/2004 S1 and R/2004 S2, particular care was taken in subtracting out background scattered light (*2*). The main Encke ringlet at 133,581 km was detected in the Voyager PPS experiment, with a measured $\tau \sim 0.10$. (The semimajor axis of Pan is 133,584 km; the ring is coincident with the orbit of the moon.) If this optical depth applied to the Cassini ISS observation, the implied fractions, by optical depth, of dust and large particles are ~10% and 90%, respectively. However, this estimate is highly uncertain, because the main Encke ringlet is known to vary with longitude.

Diffuse ring	Radius (midpoint, km)	*I/F*	τ_{min}	τ_{max}
Maxwell I	87,418	2×10^{-4}	9×10^{-5}	0.05
Huygens I	117,744	5×10^{-4}	2×10^{-4}	0.04
Encke I	133,486	0.0048	0.003	0.04
Encke	133,581	0.0268	0.016	0.21
Encke O1	133,660	2×10^{-4}	1×10^{-4}	0.002
Encke O2	133,719	0.0019	0.001	0.015
Atlas (R/2004 S1)	137,630	6×10^{-5}	3×10^{-5}	3×10^{-4}
R/2004 S2	138,900	3×10^{-6}	1×10^{-5}	2×10^{-4}

The most obvious examples of straw have length scales of a few kilometers. In Fig. 5H, the straw appearing in the broadest trough has a cant angle to the orbital motion of 11° in the trailing direction. "Light" straw has length scales typically on the order of 1 km or less, appearing most prominent closer to the wave peaks and less so in the very bottom of the trough.

Figure 6C is a dark-side SOI image of the outer edge of the Encke gap taken 18° upstream from Pan. Regularly spaced, ~6-km-wide bands bordered by the narrow dark lanes or wakes caused by Pan are visible. The ropy features seen between the first two wakes in Fig. 6C are unique in all Cassini images taken so far. They are generally between 10 and 20 km long and have a cant angle of 5° relative to the orbital direction in the trailing sense in the band closest to the gap, and ~ 9° in the next band out.

The straw and the ropy features appear similar to the transient, azimuthally limited, wakelike structures caused by gravitational instabilities in models originally used to explain the spiral arms of disk galaxies (*54*). Those structures are expected in the middle A ring and are believed to be responsible for the A-ring azimuthal brightness asymmetry. However, the observed scale in the rings is very different for the two phenomena: The gravitational wakes for the mid-A ring have length scales of ~120 m, whereas the straw is several kilometers long and the ropy structure at the outer edge of the Encke gap is tens of kilometers long. To explain the most prominent straw (in the Janus 6:5 wave) solely in terms of a gravitational instability would require a background surface mass density in the rings of $\sigma \sim 550$ g/cm^2. This is not reasonable given the optical depth of the

region. Measured surface densities in the outer A ring are an order of magnitude smaller (*42*).

These features may be caused by a combination of gravitational and kinematical effects, as found in numerical simulations of the outer edge of the Encke gap (*55*). In that work, strawlike structures were produced, even in the absence of self-gravity, when particles were forced into close proximity as their streamlines were squeezed together in the wakes created by Pan; their cant angles changed with the changing phase of the material relative to the crest of the wave. However, when self-gravity was added, they were even more pronounced and become robust structures that look like the ropy features on the SOI dark-side image of the Encke gap. We surmise that the straw and ropy features are a product of enhanced kinematical and self-gravitational effects associated with passage of ring particles, and the consequent squeezing of their streamlines, through the crests of density waves or wakes.

An unusual mottled-looking narrow region, with a radial width varying with longitude from 5 to 10 km, was observed ~60 km inside the outer edge of the A ring (Fig. 6, A and B). This region is characterized by blotchy light and dark areas about 30 to 40 km in azimuthal extent. It does not cover the whole longitudinal extent of the images in which it appears (~5°), but radially narrows and fades to nothing near the ansa. The observed extent is ~3.5°, but the actual extent could be longer if it extends beyond the left-hand part of the image.

There are similar, narrow mottled-looking regions seen in the dark-side images, but none on such a large scale as that seen at the edge of the A ring. These others are observed in the dark crests of certain waves or in the wakes caused by Pan. The two most prominent of these are the eighth wake of Pan, about 69 km exterior to the Encke gap (Fig. 6C), and the first crest of the Pandora 13:12 (Fig. 5H). These both are about 3 km wide in the radial direction and extend over most of the field of view, ~1° in longitude.

The mottled regions are also probably caused by particle clumping that is longitudinally variable. The outer A ring is shaped into a seven-lobed pattern by a 7:6 Lindblad resonance with the co-orbital satellites, and the resonant perturbations in this region are complicated by the presence of many potential components with different pattern speeds, due to the presence of two satellites, Janus and Epimetheus, within 50 km of each other in orbital distance (*37*). It is possible that these perturbations, which extend 60 km inside the A ring, may be related to the co-orbital's nonaxisymmetric perturbations in this region. The second type of mottling seen in the dark crests of waves and in the Pan wakes may be nothing more than extreme particle clumping that occurs in regions that are kinematically compressed and eventually evolves into straw and ropy structure after the particles have passed through the crests/wakes and into the troughs.

We have discovered and measured the properties of new narrow tenuous rings between the A ring and the F ring and within some of the gaps in Saturn's main rings: i.e., Encke in the A ring, Huygens in the inner Cassini division, and Maxwell in the C ring (*2*) (Fig. 7 and Table 3). Ringlets had been observed inside the Encke gap in Voyager images, but the Cassini lit-side image shows the presence of three major ones and a rather tenuous one that may be nothing more than an enhancement on a broader sheet of material. The center ringlet, which probably has the highest optical depth among the ringlets in Table 3, is coincident with Pan's orbit, implying that here, the ring particles are being maintained in horseshoe orbits (*56–58*). It, along with other ringlets in the Encke gap, exhibits longitudinal variations in brightness, suggestive of the particle accumulation that occurs in corotation sites associated with corotation resonances (*58a*).

It is not clear at present whether the origin of all these low–optical depth ringlets is the same. The association of the Atlas ring with Atlas and the main Encke ringlet with

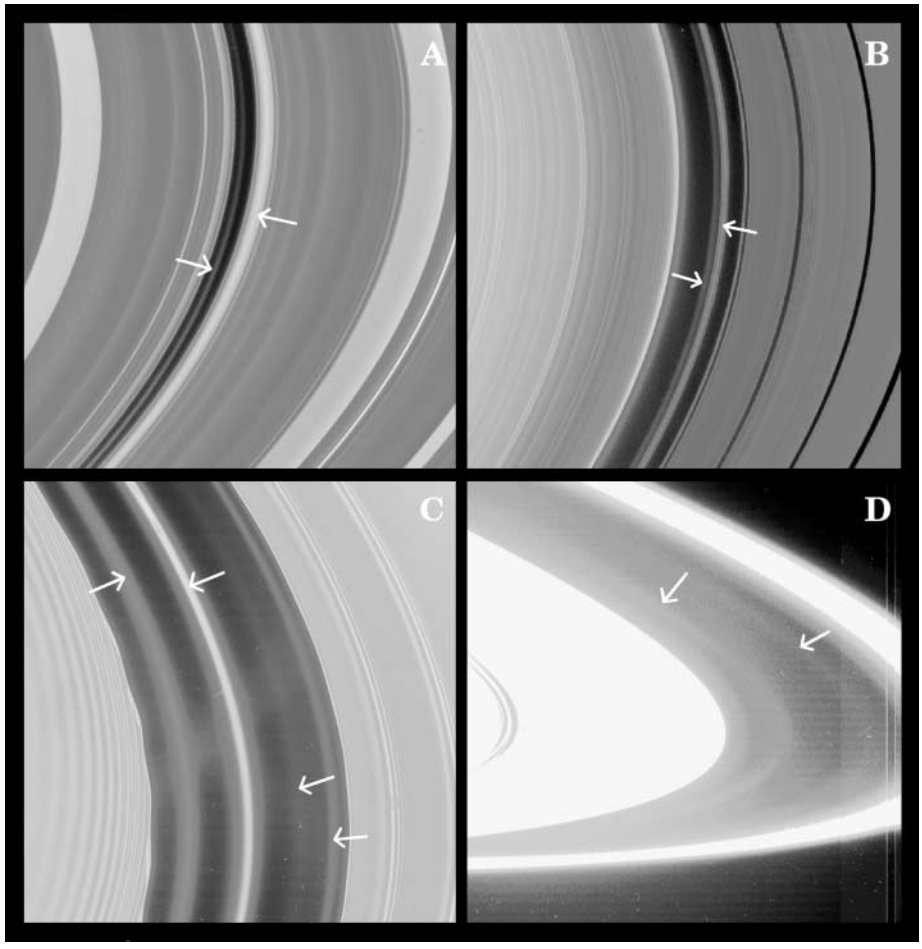


Fig. 7. Diffuse ringlets discovered within Saturn's rings. All images have been heavily processed to bring out low–optical thickness rings. Images (A) and (B) were taken with a solar incidence angle of 114.5°, an emission angle of 99°, and a phase angle of 145°. (**A**) Image N1477742794 is centered on a radius of 84,394 km, has a radial scale of 7 km/pixel, and shows a radial region of 4290-km extent, including the Maxwell gap in the C ring. The right arrow points to the optically thick Maxwell ringlet; the left arrow points to the new diffuse ring seen interior to it. (**B**) N1477740094 is centered on a radius of 117,292 km, has radial scale of 7 km/pixel, and covers 3870 km, including the Huygens gap immediately outside the outer B-ring edge. The right arrow points to the optically thick Huygens ring; the left arrow points to the new diffuse ring interior to it. Images (**C**) and (**D**) are from the dayside SOI sequence and have an incidence angle of 114.5° and an emission angle of ~94°. (C) Image N1467351325 is centered on a radius of 133,557 km interior to the Encke gap, has a radial scale of 1.15 km/pixel, a radial extent of 858 km, and was taken from a phase angle of 134°. The arrows point to (from the left) the Encke I, main Encke, Encke O1, and Encke O2 ringlets (Table 3). (D) This image is a composite of WAC images of the F-ring region taken from a phase angle of 87°. The radial scale is ~10 km/pixel. The left arrow points to R/2004 S1 (*5*) (the "Atlas" ring); the right arrow, to R/2004 S2 (*4*) (Table 3).

Pan would suggest that these rings derive from their associated moon. In other cases, a ring may exist because the material (or small parent bodies within it) are shepherded by a larger moon also present in the gap. The particles in many or all of the ringlets in Table 3 may have substantial fractions of micrometer-sized dust, implying that nongravitational forces may also affect the ringlets' dynamics. In any case, the presence of narrow diffuse ringlets in a gap like Maxwell and Huygens, along with the major Maxwell and Huygens ringlets, and the additional narrow ringlets in the Encke gap, suggests that there are other moonlets in these gaps.

Cassini images have shown clear evidence for two narrow strands of material on either side of the main core of the F ring: a 700-km-wide envelope of material surrounding the F ring, and an azimuthal structure within the ring and its envelope, some of which is clearly explained by interactions with Prometheus. Even though the F ring has been successfully modeled as an eccentric, inclined ring (*59*), its behavior in Cassini images suggests that it exhibits localized (and hence short-term) radial fluctuations, much of which is still unexplained.

A discontinuity in the ring and a streamer of material linking it with Prometheus (Fig. 8G) were observed in the vicinity of conjunction with Prometheus at its apoapse, reminiscent of the structures seen in the numerical modeling of a multistranded F ring perturbed by the moon (*60*, *61*). These models provide qualitative explanations for such phenomena: The moon's perturbation switches sign at the encounter longitude, altering ring particle semimajor axes and eccentricities (*60*). Particles encountering Prometheus immediately before apoapse are dragged backward, whereas those encountering it just after apoapse are tugged forward. The induced changes in semimajor axis remove particles from the F-ring region, opening a gap (*61*); the affected particles form a stream of material approximately in the radial direction. Physical collisions with Prometheus are not the most important effect.

Some images showed evidence for streamers extending between the F-ring core and the outer and inner strands (Fig. 8, A and B, respectively). Others (Fig. 8, C and D) have shown that gradual radial displacement in the core extending over ~12° in longitude and peaking at ~60 km in radius appeared on one ansa and was detected as a mirror image reflected in the azimuthal direction, on the other ansa 8 hours later. This suggests that these structures are due to a gradual change in eccentricity along 12° of longitude followed by a sudden change of ~0.0002, giving rise to the kink. The symmetry arises because the particles at apoapse in the ansa in Fig. 8C will be at periapse in the ansa in Fig. 8D. A corollary of this is that a kink imaged at one longitude may appear as a clump at ± 90° from that longitude.

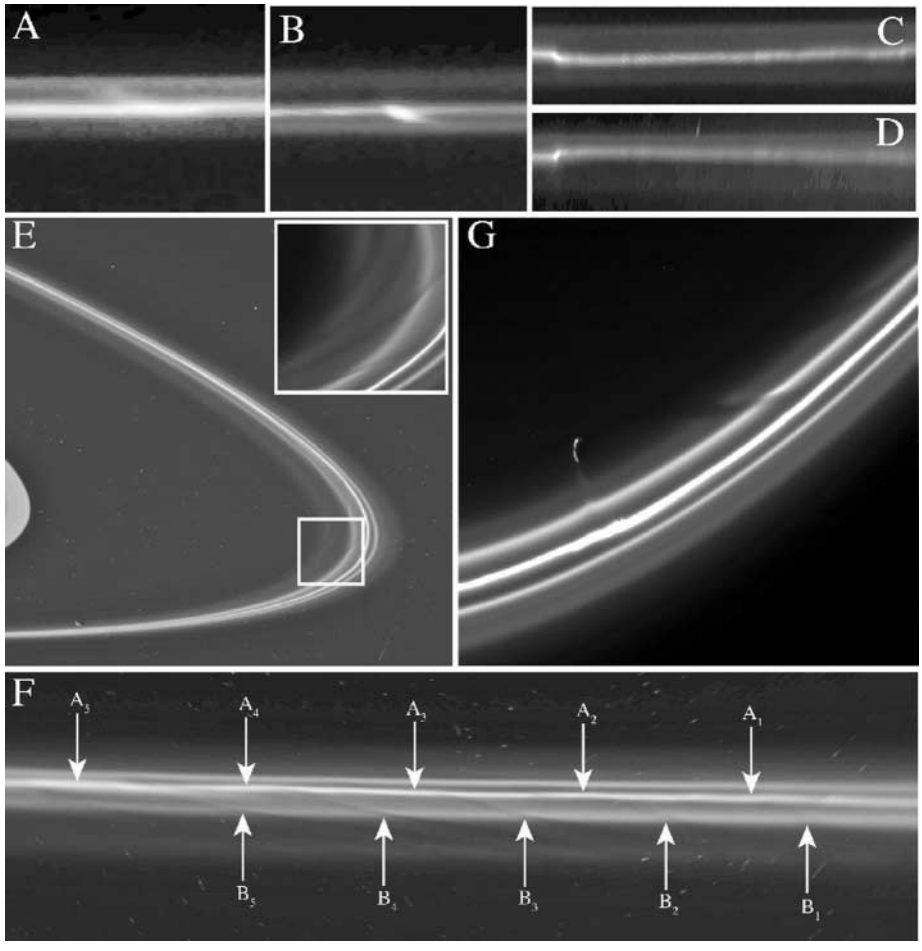


Fig. 8. (**A** and **B**) Evidence for material connecting the bright, central core of the F ring to the prominent outer and inner F-ring strands, respectively. Each reprojected image covers 6.25° of longitude and 2000 km in radius. (**C** and **D**) Images of a single reprojected section of 1000-km radial width of the F ring but separated by 8 hours and 193° in longitude. Evidence of reflective symmetry is apparent along ~12° in the azimuthal direction and in the mirrored "kink." (**E**) A view of the lit-face of the F-ring region taken on 1 July 2004 following the post-SOI descending ring plane crossing. The WAC image covers 7° in longitude at the edge of the A ring (seen on the left) and 27° at the F ring; Prometheus is visible at the lower left. The inset (top right) shows a simultaneously shuttered NAC image centered on the inner F ring. (**F**) A reprojection of the F ring region in (E) into a longitude-radius system centered on the ansa covering 18° in longitude and 1880 km in radius. The A- and B-labeled arrows denote the longitudes where Prometheus was at apoapse based on the relative mean motion of material in the core (*59*) and a strand with a semimajor axis that was 140 km interior to it, respectively. The observed slope in the reprojection is due to the eccentricity of the rings. (**G**) NAC image taken on 29 October 2004 showing the multiple strands of the F ring, a streamer of material connecting Prometheus with the F ring, and sheared gaps in the inner strands due to previous apoapse encounters with Prometheus.

Some of the clumps for which the F ring is renowned undoubtedly have their origin in this dynamic behavior caused by Prometheus.

At high resolution, two main strands are visible ~90 km exterior and ~140 km interior to the main F-ring core. Based on radial scans of the NAC, the estimated widths of the inner, core, and outer strands are 70, 20, and 30 km, respectively. All three of these rings are themselves embedded in the ~700-km-wide envelope seen in the enhanced image. Striations extending from the F-ring core, through the inner strand and into the interior dust sheet, take the appearance of "drapes" of material (Fig. 8E). The positions of the gaps in between the drapes in the inner strand are clearly correlated with the positions of Prometheus when it reaches consecutive apoapses; the positions of azimuthal features in the core of the ring are not so well correlated with Prometheus's apoapses (Fig. 8F). Furthermore, the observed changing tilt of the drapes with longitude around the ring, and asymmetry in the tilt on either side of the gap, suggest that the striations are the result of the Keplerian shear of perturbed material that was approximately radial at the time of successive apoapse passages of Prometheus, consistent with the physical model for gap formation (*60*, *61*) discussed above.

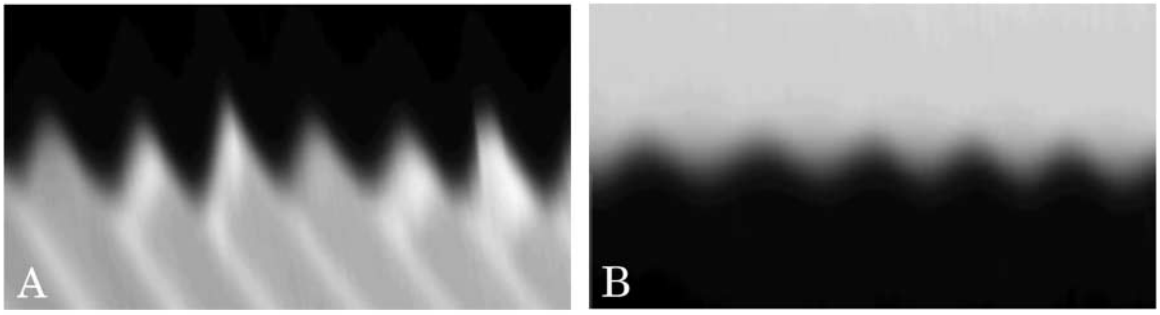


Fig. 9. Encke Gap. (**A**) Inner and (**B**) outer edges of the Encke gap as seen in Fig. 7C, mapped into a longitude-radius system, enhanced in contrast and brightness and radially stretched by a factor of 20.

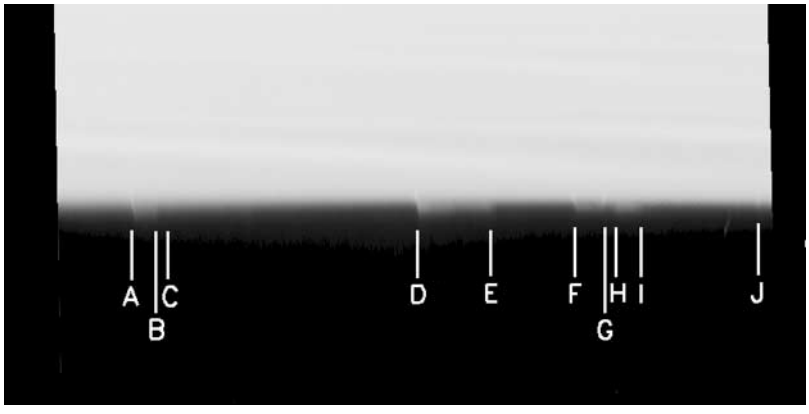


Fig. 10. Keeler gap. Spikes (sharp, radial discontinuities) and wisps (broad, faint material generally seen in between the spikes) observed in the outer edge of the Keeler gap during the SOI dayside sequence. The most easily seen features are labeled A through J. The image has been mapped into a longitude-radius system, contrast enhanced and radially stretched by a factor of 5. The longitudinal extent of the map is 4.6°; the radial extent, top to bottom, is ~60 km.

It is unlikely that Prometheus is the only perturber in the F ring. Small moons and short-lived "rubble piles" in the vicinity of the F ring have been proposed to explain a variety of Voyager (*62*–*64*) and ground-based (*6*, *65*) observations. Fourier analyses of the azimuthal structures seen in the F-ring core in Cassini movie images covering an entire F-ring orbit have confirmed the periodic perturbations of Prometheus on the ring at a wavelength of 3.2°, and reasonably strong signatures between and including 5.0° and 6°. The latter may be the signature of Pandora, which should fall at 5.75°. However, they show no evidence of any signatures of objects in the orbits of S/2004 S3, S/2004 S4, and S/2004 S6 (*2*), and with the exception of the signature of Prometheus, no signatures at the wavelengths found by previous investigators (*63*).

There is also a clear indication of material extending ~400 km beyond the edge of the A ring (Fig. 9), as well as two new diffuse rings: a ~300-km-wide ring of material, R/2004 S1, in the orbit of Atlas (*5*) and another ring, R/2004 S2 (*4*), comparable to the Atlas ring and immediately interior to Prometheus's periapse distance. These rings have *I*/*F* values comparable to those of the jovian ring (Table 3 and fig. S6). Prometheus's apoapse distance corresponds to the inner sharp boundary of the F ring's envelope. These observations indicate that Prometheus has swept material from the region occupied by its orbit.

The SOI sequence captured dark- and lit-side views of the Encke gap and the unmistakable perturbations due to Pan—scallops on the inner edge and bright spiraling streamers radiating away from the same edge. When the image is mapped in longitude and radius, severely contrast enhanced, and stretched radially by a factor of 20, it is evident that the inner edge (Fig. 9A) sports a decidedly nonsinusoidal distortion; the outer edge (Fig. 9B), a more or less sinusoidal wave of lesser amplitude.

The simple theory for the impulse approximation, in which the wavelength of the perturbation produced by a satellite on a circular orbit is given by $3\pi x$, where x is the distance of the satellite from the ring particle, does not explain all the details seen in the edges of the Encke gap. The wavelengths and amplitudes of the undulations on the inner and outer edges have been measured by least-squares fitting of a sinusoidal wave form. Fourier analysis was also used to ascertain the presence of other wavelengths, especially in the nonsinusoidal inner edge. For the inner edge, the dominant wavelength by both methods is 0.67° and $ae \sim 1.6$ km, respectively. The semimajor axis of Pan, determined by this work (Table 1), is 133584.0 ± 1 km. The locations of the gap's inner and outer edges are 133,423.5 ± 0.5 km and 133,745.1 ± 0.5 km (*66*, *67*), respectively. The distance of Pan from the inner edge is then ~160 km. This is consistent with the distance computed from the wavelength of the observed wave on the inner edge—~165 km—given the uncertainties in the latter. The kinematical explanation for the wavelength of the inner-edge wave fits the observations.

Pan's longitude at the time of the image is 19.5°. The longitude of the center of the image is 32°. The particles comprising the inner-edge waves, consequently, are leading Pan by 12.5°, which amounts to ~19 wavelengths. Thus, the inner waves are probably not damped at all, and we can estimate Pan's mass from their amplitude. The standard treatment predicts $ae \sim 2.24\ \mu a^3/x^2$, which gives $\mu = M_{\mathrm{Pan}}/M_{\mathrm{S}} \sim 8.7 \pm 1.3 \times 10^{-12}$, or $GM_{\mathrm{Pan}} = 3.3 \pm 0.5 \times 10^{-4}$ km^3 s^{-2}. This is about three times larger than the value we derive for Pan's mass from the 7:6 linear density wave. The density wave analysis is likely to be in error for reasons previously given. However, neglect of either the self-gravity of the ring particles and/or the eccentricity of Pan in the kinematic edge wave analysis that forms the basis of the relation between wave amplitude and distance may also contribute to the discrepancy. The nonsinusoidal nature of the inner-edge wave hints that Pan's eccentricity is likely to be a factor.

Edge waves are seen in the outer Encke gap edge, in which the particles are trailing Pan by 348° or ~457 wavelengths: Their dominant wavelength is 0.81° and their amplitude ae = 0.4 km. The distance of Pan from the outer edge is ~161 km, a value that is not consistent with such a wavelength (which yields $x \sim 201$ km). Either Pan is not responsible for the waves seen in the outer edge (Fig. 9B) and they indicate the presence of another moonlet in the gap ~ 40 km interior to Pan, or the simple kinematical model for edge wave development by Pan does not apply to this edge. The second putative moonlet would have to be ~25% of Pan's mass, and for the same density, ~60% of Pan's size. Such a moon would have been easily detected in Cassini movie sequences of the Encke gap. On the other hand, it is hard to imagine a pro-

cess that would alter the wavelength/distance relation on the outer edge and not on the inner edge, because the ring material interior and exterior to Pan has comparable optical depths and therefore surface mass densities. Cassini images of the Encke gap, and the large number of wakes due to Pan and second- and third-order waves, suggest that there is far less damping than the simplest, streamline intersection treatment of this dynamical system would predict (*52*, *68*, *69*).

The Keeler gap, the narrow gap only ~250 km interior to the outer edge of the A ring, was also captured in the lit- and dark-side SOI image sequences. In Cassini images, the gap is ~42 km wide. Several faint discontinuities, or spikes, in the outer gap edge have been discovered in two NAC dayside images (Fig. 10); these are similar to the spikes protruding inward from the core of the F ring during Prometheus's passages (Fig. 8G). Wisps of faint material, vaguely reminiscent of the drapes and striations seen interior to the F ring, are seen in association with, and between, some of these spikes in the Keeler gap outer edge (Fig. 10). These features all move in unison at the Keplerian rate appropriate to the motion of particles at this location. Arguing on the basis of their similarity to the features caused by Prometheus in the F ring, it is likely that they are caused by the apoapse passages of a yet-unseen moonlet on an eccentric orbit within the Keeler gap. The spikes are ~5 km long. The putative moon would have a diameter of a few kilometers, scaling from the widths of the Encke and Keeler gaps and our inferred size for Pan (*28*). Interpreted in this way, measurement of their preferred length scale should indicate the position of the moonlet. To this end, correlations between the spikes' longitudinal positions and grids with a range of characteristic spacings were numerically calculated. The most statistically significant characteristic spacing was found to be 199 km at a 78% confidence level. A set of features with this particular wavelength, caused by a moon on an eccentric orbit, would be associated (through the $3\pi x$ relation) with a perturber at a radial separation of 21 km, which is half the width of the Keeler gap. These observations indicate that there is a tiny moon orbiting at the center of the Keeler gap.

Finding any feature in the outer part of the A ring that is certifiably circular is difficult; almost all features in this region are noncircular density waves or nonaxisymmetric edges. Taking the outer edge of the Keeler gap to be circular (despite the presence of the short spatial scale spikes and wisps), streamer-like features exterior to the outer Keeler gap edge are then spiral in nature, like those seen spreading away from the inside of the Encke gap edge. These may be wake structures caused by the putative Keeler moon.

References and Notes

1. C. C. Porco *et al.*, *Space Sci. Rev.* **115**, 363 (2004).
2. See supporting data in *Science* Online.
3. C. C. Porco *et al.*, *IAU Circ.* **8389**, 1 (2004).
4. C. C. Porco *et al.*, *IAU Circ.* **8432**, 1 (2004).
5. C. C. Porco *et al.*, *IAU Circ.* **8401**, 1 (2004).
6. J. M. Barbara, L. W. Esposito, *Icarus* **160**, 161 (2002).
7. P. Thomas, J. Veverka, D. Morrison, M. Davies, T. V. Johnson, *J. Geophys. Res.* **88**, 8743 (1983).
8. N. Borderies, P. Goldreich, S. Tremaine, in *Planetary Rings*, R. Greenberg, A. Brahic, Eds. (Univ. of Arizona Press, Tucson, AZ, 1984), pp. 713–734.
9. J. Scargle *et al.*, *Bull. Am. Astron. Soc.* **25**, 1103 (1993).
10. J. A. Burns, B. J. Gladman, *Planet. Space Sci.* **46**, 1401 (1998).
11. R. A. Jacobson, *Bull. Am. Astron. Soc.* **36**, 1097 (2004).
12. P. Goldreich, R. Sari, *Astrophys. J.* **585**, 1024 (2003).
13. R. A. Jacobson, R. G. French, *Icarus* **172**, 382 (2004).
14. J. N. Cuzzi, R. G. French, L. Dones, *Icarus* **158**, 199 (2002).
15. The reflectivity of a surface is given in terms of *I*/*F*, where *I* is the intensity of the scattered light and πF is the net flux of solar radiation per unit area normal to itself (*70*). Spectral slope is defined as $d(I/F)/d\lambda \div (I/F)_0$, where $(I/F)_0$ is the value of *I*/*F* at some fiducial wavelength. Ground-based observations have shown that the main rings are "red" in the spectral range from 300 to ~600 nm (i.e., *I*/*F* increases with wavelength, λ), whereas *I*/*F* is roughly constant from 600 to 1000 nm (*71*, *72*). HST observations showed that the rings' spectral slope increases with phase angle over the range 0° to 6° visible from Earth. This reddening is thought to result from multiple scattering within the surfaces of ring particles (*14*). Our results show that little, if any, additional "phase reddening" occurs at phase angles between 6° and 66°. Color variations in Saturn's rings on a variety of spatial scales have been seen by Voyager and HST (*16*, *73*, *74*).
16. J. N. Cuzzi *et al.*, in *Planetary Rings*, R. Greenberg, A. Brahic, Eds. (Univ. of Arizona Press, Tucson, AZ, 1984), pp. 73–199.
17. F. Poulet, J. N. Cuzzi, *Icarus* **160**, 350 (2002).
18. F. Poulet, D. Cruikshank, J. N. Cuzzi, T. L. Roush, R. G. French, *Astron. Astrophys.* **412**, 305 (2003).
19. L. R. Doyle, L. Dones, J. N. Cuzzi, *Icarus* **80**, 104 (1989).
20. J. N. Cuzzi, P. R. Estrada, *Icarus* **132**, 1 (1998).
21. B. A. Smith *et al.*, *Science* **212**, 163 (1981).
22. C. A. McGhee *et al.*, *Icarus,* in press.
23. T. Nitter, O. Havnes, F. Melandsø, *J. Geophys. Res.* **103**, 6605 (1998).
24. The interpretation of dark-side images can be ambiguous: Empty space, as well as high–optical depth regions, can appear; intermediate–optical depth regions appear bright.
25. C. C. Porco *et al.*, *Icarus* **60**, 1 (1984).
26. C. C. Porco, *Adv. Space Res.* **10**, 221 (1990).
27. Density waves are often referred to by the notation $(m + n + p)$:$(m - 1)$, where *m*, *n*, and *p* are integers. This ratio is approximately equal to the ratio of the mean motions of the ring particle and the satellite that produces the wave. For a linear wave, the amplitude is proportional to $M_s e^p (\sin I)^n$, where M_s, *e*, and *i* are the mass, orbital eccentricity, and inclination to Saturn's equator of the satellite, respectively. For a given moon, the strongest density waves have $p = n = 0$, so that these waves are of the form m:$(m - 1)$. Waves driven by interior moons on exterior rings, like the Uranian ε ring, would have $(m - 1)$:m for first-order waves, and $(m - 2)$:m for second-order waves. The waves we describe in detail are Atlas and Pan waves with $m = 5$ and 7, respectively.
28. P. Goldreich, C. C. Porco, *Astron. J.* **93**, 730 (1987).
29. J. C. B. Papaloizou, D. N. C. Lin, *Astrophys. J.* **331**, 838 (1988).
30. C. C. Porco, thesis, California Institute of Technology (1983).
31. E. Turtle, C. C. Porco, V. Haemmerle, W. Hubbard, R. Clark, *Bull. Am. Astron. Soc.* **23**, 1179 (1991).
32. E. I. Chiang, P. Goldreich, *Astrophys. J.* **540**, 1084 (2000).
33. L. J. Horn, J. N. Cuzzi, *Icarus* **119**, 285 (1996).
34. S. Tremaine, *Astron. J.* **125**, 894 (2003).
35. E. Griv, M. Gedalin, *Planet. Space Sci.* **51**, 899 (2003).
36. J. Schmidt, H. Salo, *Phys. Rev. Lett.* **90**, 061102 (2003).
37. C. Porco, G. E. Danielson, P. Goldreich, J. B. Holberg, A. L. Lane, *Icarus* **60**, 17 (1984).
38. F. Namouni, C. Porco, *Bull. Am. Astron. Soc.* **34**, 884 (2002).
39. L. W. Esposito, M. O'Callaghan, R. A. West, *Icarus* **56**, 439 (1983).
40. F. H. Shu, in *Planetary Rings*, R. Greenberg, A. Brahic, Eds. (Univ. of Arizona Press, Tucson, AZ, 1984), pp. 513–561.
41. P. A. Rosen, J. J. Lissauer, *Science* **241**, 690 (1988).
42. L. J. Spilker *et al.*, *Icarus* **171**, 372 (2004).
43. H. A. Zebker, E. A. Marouf, G. L. Tyler, *Icarus* **64**, 531 (1985).
44. R. G. French, P. D. Nicholson, *Icarus* **145**, 502 (2000).
45. P. Goldreich, S. D. Tremaine, *Icarus* **34**, 227 (1978).
46. P. D. Nicholson, D. P. Hamilton, K. Matthews, C. F. Yoder, *Icarus* **100**, 464 (1992).
47. C. A. McGhee, P. D. Nicholson, R. G. French, K. J. Hall, *Icarus* **152**, 282 (2001).
48. S. Renner, B. Sicardy, R. G. French, *Icarus,* in press.
49. F. H. Shu, L. Dones, J. J. Lissauer, C. Yuan, J. N. Cuzzi, *Astrophys. J.* **299**, 542 (1985).
50. N. Borderies, P. Goldreich, S. Tremaine, *Icarus* **63**, 406 (1985).
51. N. Borderies, P. Goldreich, S. Tremaine, *Icarus* **68**, 522 (1986).
52. M. R. Showalter, J. N. Cuzzi, E. A. Marouf, L. W. Esposito, *Icarus* **66**, 297 (1986).
53. L. J. Horn, M. R. Showalter, C. T. Russell, *Icarus* **124**, 663 (1996).
54. W. H. Julian, A. Toomre, *Astrophys. J.* **146**, 810 (1966).
55. M. C. Lewis, G. R. Stewart, *Icarus,* in press.
56. S. F. Dermott, T. Gold, A. T. Sinclair, *Astron. J.* **84**, 1225 (1979).
57. S. F. Dermott, C. D. Murray, A. T. Sinclair, *Nature* **284**, 309 (1980).
58. S. F. Dermott, C. D. Murray, *Icarus* **48**, 1 (1980).

58a. C. C. Porco, *Science* **253**, 995 (1991).

59. A. S. Bosh, C. B. Olkin, R. G. French, P. D. Nicholson, *Icarus* **157**, 57 (2002).
60. M. R. Showalter, J. A. Burns, *Icarus* **52**, 526 (1982).
61. S. M. Giuliatti Winter, C. D. Murray, M. K. Gordon, *Planet. Space Sci.* **48**, 817 (2000).
62. J. N. Cuzzi, J. A. Burns, *Icarus* **74**, 284 (1988).
63. R. A. Kolvoord, J. A. Burns, M. R. Showalter, *Nature* **345**, 695 (1990).
64. C. D. Murray, M. K. Gordon, S. M. Giuliatti Winter, *Icarus* **129**, 304 (1997).
65. F. Poulet, B. Sicardy, P. D. Nicholson, E. Karkoschka, J. Caldwell, *Icarus* **144**, 135 (2000).
66. P. D. Nicholson, M. L. Cooke, E. Pelton, *Astron. J.* **100**, 1339 (1990).
67. R. G. French *et al.*, *Icarus* **103**, 163 (1993).
68. J. N. Cuzzi, J. D. Scargle, *Astrophys. J.* **292**, 276 (1985).
69. N. Borderies, P. Goldreich, S. Tremaine, *Icarus* **80**, 344 (1989).
70. S. Chandrasekhar, *Astrophys. J.* **106**, 152 (1947).
71. K. Lumme, H. J. Reitsema, *Icarus* **33**, 288 (1978).
72. E. Karkoschka, *Icarus* **111**, 174 (1994).
73. P. R. Estrada, J. N. Cuzzi, *Icarus* **122**, 251 (1996).
74. P. R. Estrada, J. N. Cuzzi, M. R. Showalter, *Icarus* **166**, 212 (2003).
75. G. W. Null, E. L. Lau, E. D. Biller, J. D. Anderson, *Astron. J.* **86**, 456 (1981).

76. We acknowledge the many individuals across the imaging team who have assisted in the design of imaging sequences and camera commands and in other vital operational and image-processing tasks, in particular E. Birath, J. Riley, B. Knowles, C. Clark, M. Belanger, and D. Wilson. M. Hedman is acknowledged for fruitful discussions. This work has been funded by NASA/JPL, the UK Particle Physics and Astronomy Research Council, the German Aerospace Center (DLR), and Université Paris VII Denis Diderot, Commissariat à l'Energie Atomique, Astrophysique Interactions Multiéchelles, France.

Supporting Online Material

www.sciencemag.org/cgi/content/full/307/5713/1226/DC1
SOM Text
Table S1
Figs. S1 to S6
References

30 November 2004; accepted 12 January 2005
10.1126/science.1108056

Cassini Imaging Science: Initial Results on Phoebe and Iapetus

C. C. Porco,[1]* E. Baker,[1] J. Barbara,[2] K. Beurle,[3] A. Brahic,[4] J. A. Burns,[5] S. Charnoz,[4] N. Cooper,[3] D. D. Dawson,[6] A. D. Del Genio,[2] T. Denk,[7] L. Dones,[8] U. Dyudina,[9] M. W. Evans,[3] B. Giese,[10] K. Grazier,[11] P. Helfenstein,[5] A. P. Ingersoll,[9] R. A. Jacobson,[11] T. V. Johnson,[11] A. McEwen,[6] C. D. Murray,[3] G. Neukum,[7] W. M. Owen,[11] J. Perry,[6] T. Roatsch,[10] J. Spitale,[1] S. Squyres,[5] P. C. Thomas,[5] M. Tiscareno,[5] E. Turtle,[6] A. R. Vasavada,[11] J. Veverka,[5] R. Wagner,[10] R. West[11]

The Cassini Imaging Science Subsystem acquired high-resolution imaging data on the outer Saturnian moon, Phoebe, during Cassini's close flyby on 11 June 2004 and on Iapetus during a flyby on 31 December 2004. Phoebe has a heavily cratered and ancient surface, shows evidence of ice near the surface, has distinct layering of different materials, and has a mean density that is indicative of an ice-rock mixture. Iapetus's dark leading side (Cassini Regio) is ancient, heavily cratered terrain bisected by an equatorial ridge system that reaches 20 kilometers relief. Local albedo variations within and bordering Cassini Regio suggest mass wasting of ballistically deposited material, the origin of which remains unknown.

Phoebe orbits Saturn at a distance of 12.9 × 10^6 km every 550 days in a retrograde orbit, which suggests that it was captured early in solar system history (*1*). The Cassini spacecraft flew within 2071 km of Phoebe to enable high-resolution remote-sensing observations and determination of the satellite's mass. Ground-based and low-resolution Voyager images had shown the satellite to be dark (normal reflectance ~0.08), with relatively small regions that were ~50% brighter (*2*). This dark surface has been hypothesized to be a source for material coating Iapetus, although it is known that Phoebe's surface is spectrally different from the dark material on Iapetus (*3*, *4*).

The Imaging Science Subsystem (ISS) (*5*) obtained images of Phoebe at better than 2 km pixel^{-1} (Fig. 1) over slightly more than three Phoebe rotations (period of 9.27 hours) (*6*); the closest approach occurred on 11 June 2004. The highest resolution images have a pixel scale of 13 m. We derived a spin-pole orientation of right ascension = 356.6°, declination = 77.9° from 211 control points measured in 80 images. The previous solution from Voyager data was only 9° different (*7*). Additional stereo solutions and matching limb and shadow positions provide a 2° by 2° digital model of the shape (*8*), from which the satellite's volume and major topographic forms can be measured. The mean radius is 106.6 ± 1 km. Phoebe's topography, relative to an equipotential surface, is within the range of other small objects and is much higher than that for clearly relaxed objects.

The calculated volume, 5.07 ± 0.14 × 10^6 km^3, combined with the mass determined from tracking the spacecraft (*9*), gives a mean density of 1630 ± 45 kg m^{-3}. The surface gravity on Phoebe, including rotational effects, ranges from 3.8 to 5.0 cm s^{-2}, and escape velocity, for objects launched vertically, ranges from 89 to 108 m s^{-1}.

Mean density is a function of both the sample density and the body's overall porosity. For porosities <~40%, Phoebe's mean density of 1630 kg m^{-3} requires a mixture of ice and rock of some type (Fig. 2). For porosities >~40%, Phoebe could be a completely rocky body. However, the identification of water ice on its surface telescopically (*10*, *11*) and evidence from the Cassini Visual and Infrared Mapping Spectrometer (VIMS) experiment (*12*) that shows widespread water ice, water of hydration, and trapped CO_2 on the surface suggest a more volatile overall composition. In addition, ISS images suggest icy materials at and near the surface. Furthermore, most examples of high inferred porosities, such as

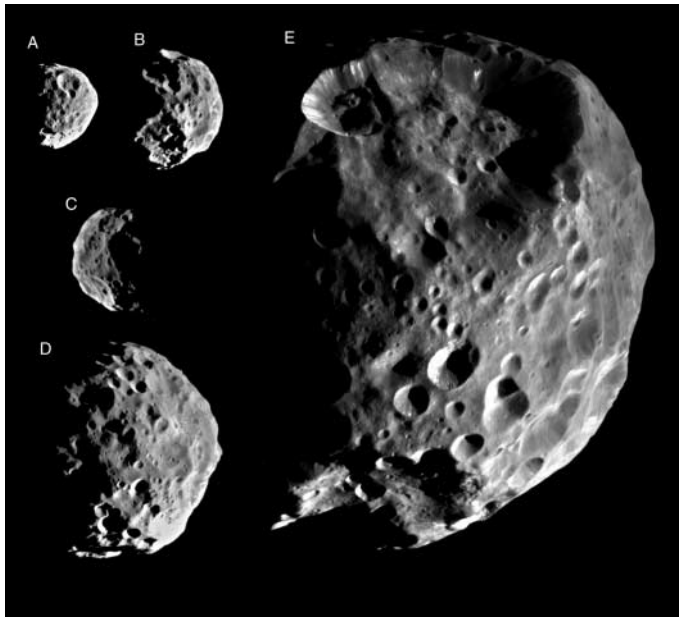


Fig. 1. Global views of Phoebe. (**A**) Image N1465644359, from 195,000 km, subspacecraft (S/C) longitude of 78°W. (**B**) N1465652463, 144,000 km, 165°W. (**C**) N1465700301, 160,000 km, 142°W. (**D**) N1465662798, 78,000 km, 276°W. (**E**) Mosaic of N1465669778 and N1465669953, 33,000 km, 349°W. Part of N1465669953 has been reprojected to match the limb of the other image. North is approximately up in all images.

[1]Cassini Imaging Central Laboratory for Operations, Space Science Institute, 4750 Walnut Street, Suite 205, Boulder, CO 80301, USA. [2]Goddard Institute for Space Studies, NASA, 2880 Broadway, New York, NY 10025, USA. [3]Astronomy Unit, Queen Mary, University of London, London, E1 4NS, UK. [4]Centre d'Etudes de Saclay, Université of Paris 7, L'Orme des Merisiers, 91191 Gif-sur-Yvette Cedex, France. [5]Department of Astronomy, Cornell University, Space Sciences Building, Ithaca, NY 14853, USA. [6]Department of Planetary Sciences, University of Arizona, 1629 East University Boulevard, Tucson, AZ 85721, USA. [7]Institut für Geologische Wissenschaften, Freie Universität, 12249 Berlin, Germany. [8]Department of Space Sciences, Southwest Research Institute, 1050 Walnut Street, Suite 400, Boulder, CO 80302, USA. [9]Division of Geological and Planetary Sciences, California Institute of Technology, 150-21, Pasadena, CA 91125, USA. [10]Institute of Planetary Research, German Aerospace Center, Rutherfordstrasse 2, 12489 Berlin, Germany. [11]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA.

*To whom correspondence should be addressed. E-mail: carolyn@ciclops.org

the co-orbital moons, Janus and Epimetheus, and the F-ring shepherds, Pandora and Prometheus, occur for objects whose central pressures are exceeded below a few kilometers depth in Phoebe. Thus, Phoebe might not sustain such porosities over a large fraction of its volume.

For zero porosity, Phoebe's density would imply a mixture of ice and rock similar to those of the large icy satellites Ganymede, Callisto, and Titan (Fig. 2). Estimates of rock fractions for the smaller icy satellites (Mimas, Tethys, Dione, Rhea, and Iapetus) range from 0.25 to 0.5 (*13*), resulting in a range of sample densities from ~1100 to 1500 kg m^{-3}. Given that some porosity is plausible for a small satellite, Phoebe's sample density is thus probably greater than that found in the icy satellites of either Jupiter or Saturn but less than the mean densities of Pluto and Triton, objects thought to be representative of Kuiper Belt Objects, with implied rock fractions of about 0.7, or an equivalent sample density of ~1900 kg m^{-3} (*14*).

Phoebe's surface is dominated by impact craters, large landslides, and local albedo markings (Figs. 1 and 3). Phoebe's craters range in diameter from the lower limit imposed by the image resolution, ~50 m, to almost the maximum limit imposed by the size of Phoebe itself, ~100 km; there are >130 craters with diameters >10 km. The vast majority of Phoebe's craters have morphologies typical of simple craters seen throughout the solar system, including the usual range of forms apparently induced by subsequent mass movement or other impacts. Because of the low gravity on Phoebe, there should be no central-peaked or multiringed craters, and this is indeed the case.

Two craters exhibit large mass-movement accumulations (Figs. 1E and 3B). Hummocky debris on the floor of the 100-km crater named Jason (Fig. 4) probably originated from the north and east slopes, which are steep sided and exhibit evidence of recent downslope motion, including some bright streamers. Crenulations in the rim of a smaller crater named Erginus correlate with lobes of debris on the floor (Fig. 1E). The larger mass-movement accumulation in Jason has sections ~2 km thick, and its total volume is estimated to be >10^3 km^3. The current height of the eastern wall ranges from 11 to 15 km, with slopes up to ~40°. This wall could have supplied the estimated volume of slumped material with less than 2 km of horizontal retreat. The deposit within Jason may have been emplaced in several events, and the debris inside crater Erginus, which postdates some of the Jason landslides, may also consist of two or more discrete flows, one of which appears to have overtopped the crater's south rim.

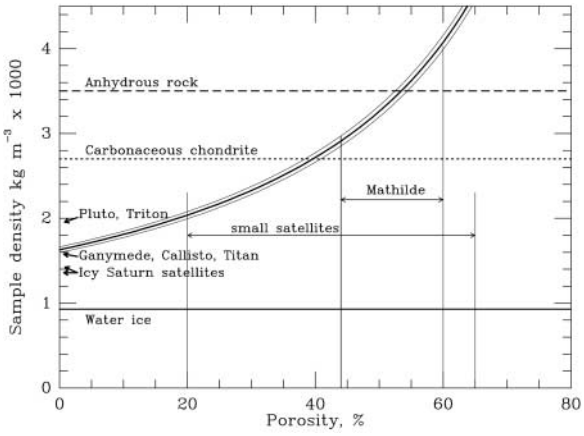


Fig. 2. Sample density for Phoebe as a function of porosity. For any given density, greater porosity requires a higher sample density. For Phoebe's mean density of 1630 kg m^{-3}, porosities less than 40% require some mixture of rock and ice; higher porosities could exist without an ice component. For comparison, calculated porosities of small, icy Saturn satellites range between 20% and 65%; asteroid Mathilde, a rocky object, has porosities as high as 60%.

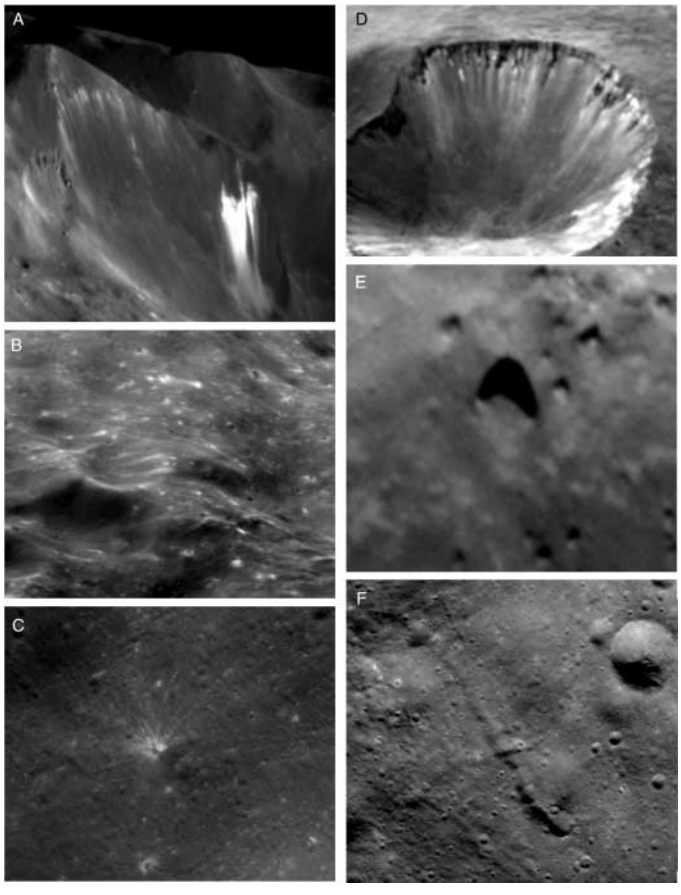


Fig. 3. Details of morphology of Phoebe. (**A**) Bright exposures and downslope streamers in Jason crater. Portion of N1465673138; image width 15 km; centered at 31°N, 302°W. (**B**) Bright markings surrounding small craters in hummocky material that probably was emplaced as landslides from the walls of crater Jason. Portion of N1465677447; image width 26 km; centered at 21°N, 318°W. (**C**) Crater with bright rays, and other craters exposing bright material. Portion of N1465674782; image width 8.2 km; centered at 1°N, 346°W. (**D**) Crater with banding in upper slopes. Portion of N1465672905; image width 17 km; centered at 36°S, 334°W. (**E**) Ejecta blocks in bottom of crater; largest is ~300 m across. Portion of N1465674693; image width 1.8 km; centered at 1°S, 353°W. (**F**) Elongate depression, possibly secondary crater chain. Portion of N1465674502; image width 10.5 km; centered at 0°N, 11°W.

Phoebe is a densely cratered object. Cumulative numbers of craters, between 100 m and 100 km diameter, per unit area, define a steep-sloped curve (Fig. 5). Crater densities approach those seen on other heavily cratered objects (*15*, *16*). This impact record may show the effects of impacts by other small, irregular saturnian satellites. Phoebe itself is the largest member of a group of at least five objects with similar inclinations and eccentricities (*17*, *18*). The population of irregular satellites should produce craters a few tens of kilometers in diameter (*17*, *19*). The high crater frequencies strongly suggest a surface at least 4 Gy old, consistent with early capture into Saturn orbit (*1*), when the outer solar system probably had a large population of small objects (*20*).

Photometric properties can help distinguish materials and geologic units. Cassini's encounter with Phoebe provided coverage in

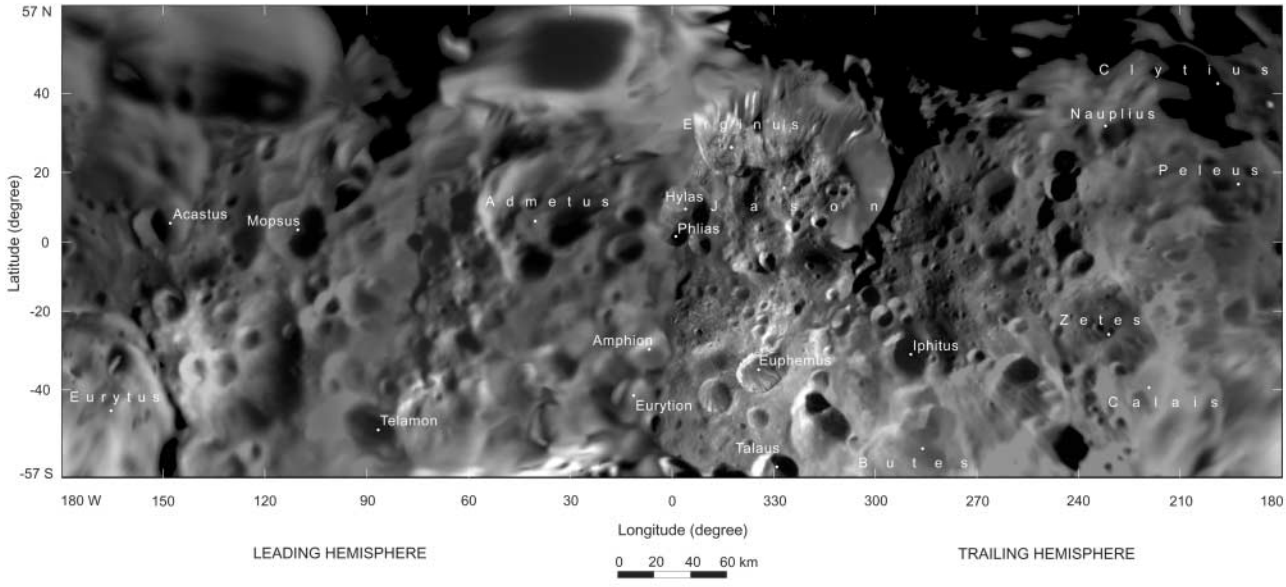


Fig. 4. A mosaic of images of Phoebe, including the provisional names assigned to some craters by the International Astronomical Union. South polar regions were imaged only at high emission angles; high northern latitudes were in darkness. Resolution and solar phase changed during flyby.

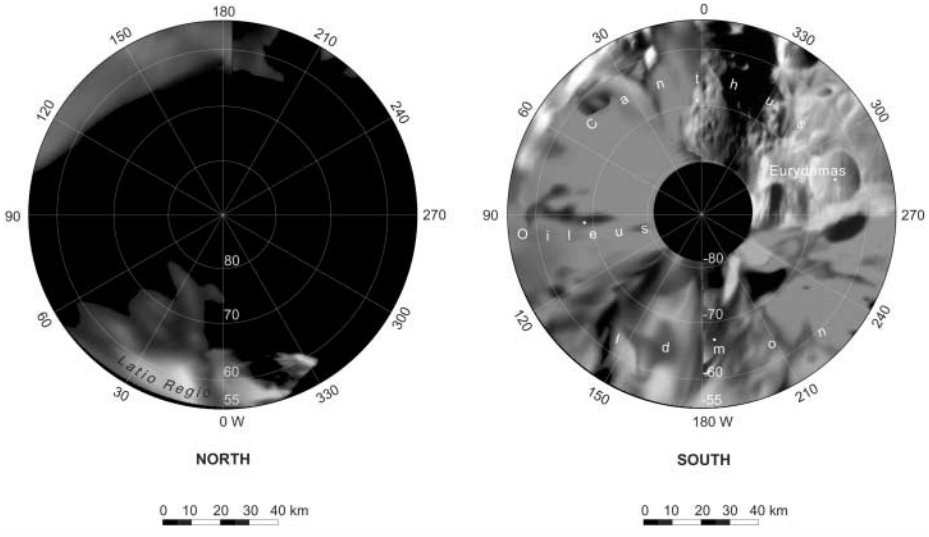
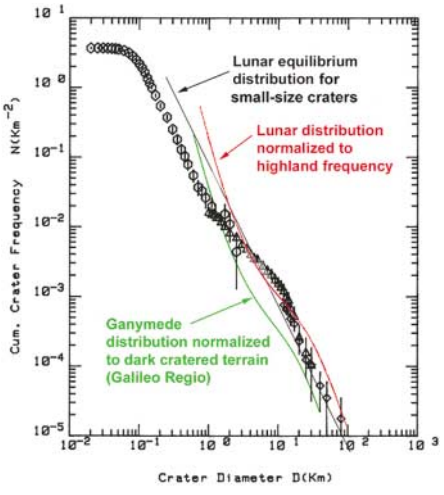


Fig. 5. Cumulative crater size-frequency distributions of Phoebe measured at image pixel scales from 4 km down to 30 m. The red curve is the lunar production function normalized to (average) lunar highland frequencies [see, e.g., (*35*)]. The green curve represents the production function derived for Jupiter's largest satellite, Ganymede, normalized to old, dark, cratered terrain of Galileo Regio (*16*, *36*). The shapes of the lunar and Ganymede production-function curve and the Phoebe crater size-frequency distributions at larger sizes all show departures from simple log functions. Also shown is the lunar equilibrium distribution with a –2 slope for small craters in the lunar maria (e.g., *16*).



solar phase angle (α) from 24° ≤ α ≤ 92°. Rotational light curves derived from the preflyby and postflyby data displayed no substantial variations of whole-disk color with longitude. At α near 90°, the average whole-disk brightness of Phoebe was about one magnitude brighter than expected by extrapolating the best-available Voyager photometric model (*2*). This discrepancy suggests that average regolith particles on Phoebe's surface scatter light more isotropically than predicted from the Voyager model, which was constrained only by data at α ≤ 33°.

We find a range of normal reflectances (r_n) of $0.07 \leq r_n \leq 0.3$. Local albedo variations on the surface of Phoebe having contrast factors of 2.0 to 3.0 as measured by ISS are manifested chiefly as brighter downslope streamers and bright annuli, rays, or irregular bright areas around small craters. Small parts of a few downslope streamers (Fig. 3A) are saturated in ISS images and thus have yet higher reflectances, but these represent a small fraction of the bright albedo features on Phoebe. These contrast ratios suggest normal reflectances of ~30% or less, values incom-

patible with clean ice. Thus, although most of the brighter outcrops are probably rich in ice, they are "dirty" (contaminant fraction could still be small) and could evolve to darker lag deposits that mantle Phoebe's surface through sublimation and thermal degradation processes related to insolation, sputtering, and impact cratering.

Bright material appears to be exposed by cratering on flat areas and gentle slopes and by mass wasting of steep scarps. Bright spots are associated with craters (Fig. 3, B and C) ranging from below the image resolution to ~1 km in size. Material excavated by impacts typically comes from depths <0.1 crater diameter (*18*); thus, the bright crater deposits represent material from a few meters to ~100 m in depth. Bright exposures also occur in landslide debris, which represents a mixture of materials from a variety of depths. Therefore, the brighter, ice-rich material occurs at shallow depths (or deeper) in widespread geographic and geologic settings. However, only a small fraction of craters, (<10%) in a limited size range (diameters <~1 km) presently displays bright materials. This observation suggests that bright materials darken or are covered as they age by processes such as infall of dark material from impacts among other small, outer satellites (*21*); deposition (regional or global) of debris excavated from elsewhere on Phoebe; sublimation of ice from the bright component; or, possibly, photochemical darkening of impurities in the brighter material.

The walls of some craters ranging from ~20 to 100 km in diameter (Fig. 3D) show evidence for layering or banding of materials at depth. In Jason (Fig. 3A), alignment of exposures of bright material suggests some layering at depths >1 km. The event that created Euphemus (Fig. 3D) may have excavated a layered surface, or the banding may represent exposure of the crater's own ejecta by slumping of oversteepened wall material near the rim.

Elongate depressions, typically less than 200 m wide and of varying continuity, are visible in some of the higher resolution images (Fig. 3F). Most of these are radially aligned with the overlapping craters Phlias and Hylas, near 0°W. The association suggests that these may be secondary craters, or fractures, or both, formed by the creation of a relatively young crater.

The presence of boulders 30 to 300 m across (Fig. 3E), collected in the bottoms of craters and scattered between craters, is fully

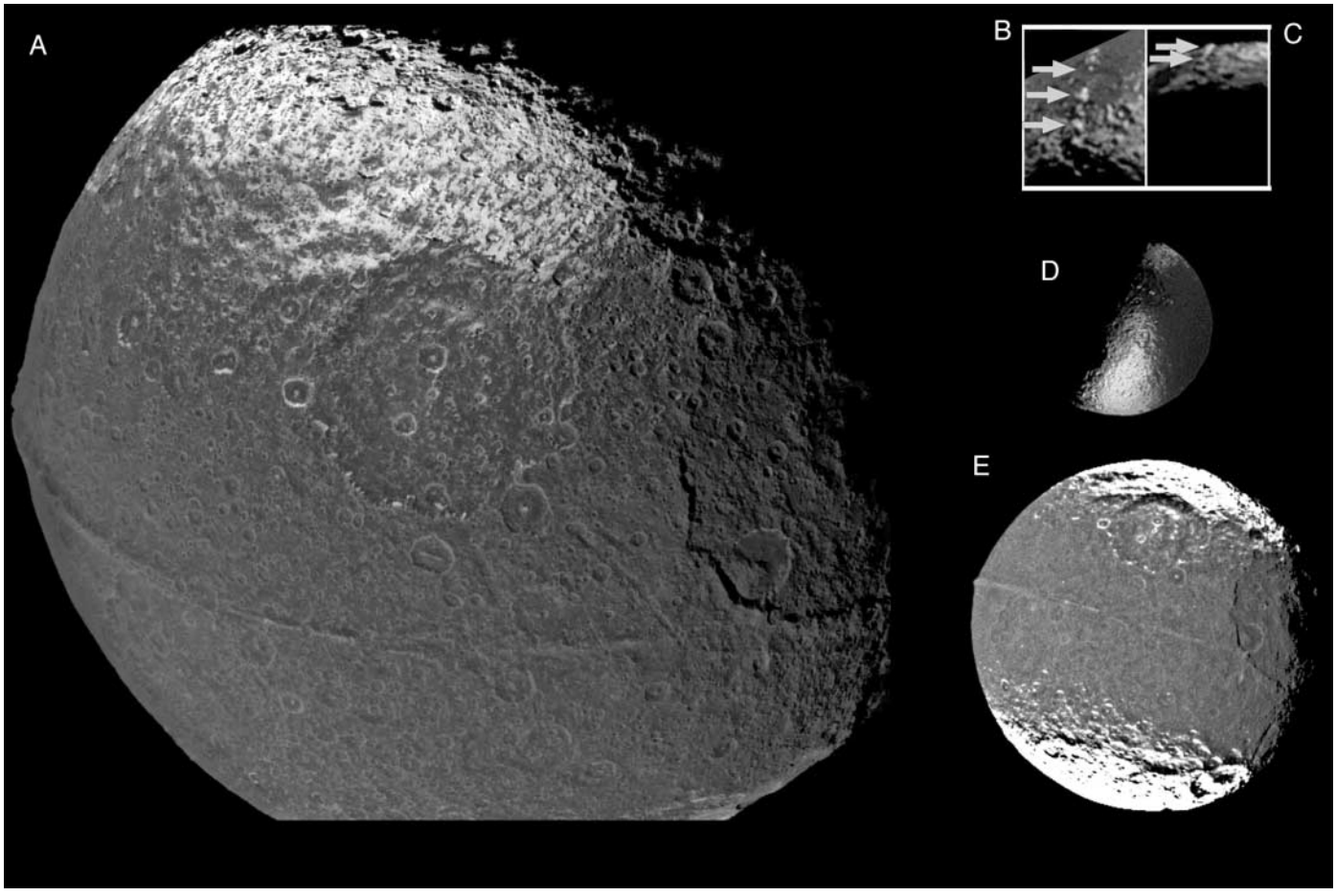


Fig. 6. Iapetus features. (**A**) Mosaic of images N1483151512, N1483152862, N1483152937, and N1483153026. The disk's middle is at 20°N, 80°W. The ridge runs roughly along 2°S and in this view traverses longitudes 40°W to 150°W, a distance of about 1400 km. Streaked dark markings in the brighter area at the top point back to the center of the leading side of Iapetus. Bright markings on the southern rim of the 400-km crater near the center indicate downslope motion and preferential exposure on north-facing slopes. (**B**) Detail of peaks along the equatorial ridge. Portion of image N1476735994, rotated; image edge at upper left; width is ~350 km. Peak near middle is at 1°S, 190°W. (**C**) Peaks seen on limb in image N1476993421, S/C = 7°S, 292°W. Peaks on limb are at ~205°W, essentially on the equator. North is to left. (**D**) View of linear chain of peaks at transition from Cassini Regio (right) to brighter terrain. Image N1476575655, S/C = 27°S, 216°W. Line of bright peaks extends from 185°W to 210°W and is within 2° of the equator. North is to upper right. Diameter of disk is ~1490 km. (**E**) Cassini Regio, bright polar areas, large basins, and equatorial ridge. Image N1482859934, S/C = 2°S, 70°W. North is to upper right. Basin on the right is ~550 km across; the one near the upper boundary of the dark area is 400 km across. Both basins show central raised structures, perhaps broad peak rings. A basin near the left limb (below equatorial ridge) is ~380 km diameter. Image has been contrast enhanced to show the features in the darker areas.

consistent with ejecta from cratering events. A 300-m block could have been produced by a ~25-km diameter crater (*22*). About 20 craters on Phoebe are between 25 and 50 km across.

Iapetus, Saturn's third-largest moon, orbits at a mean distance of 3.56×10^6 km on a slightly inclined prograde orbit. Cassini ISS obtained images of Iapetus during its first three orbits and during a flyby on 31 December 2004 with pixel scales that improve substantially on Voyager coverage. The most enigmatic feature of Iapetus from previous observations is the dark, roughly elliptical region centered on the leading side ("Cassini Regio") (*23*), darker by a factor of about 10 than parts of the trailing side (*24*). Theories for the origin of this feature have primarily involved the effects of impacts of material (exogenic theories), either by direct coating of the surface or by erosional effects of impacts (*3*, *25–29*). The source of possible impactors—whether the remnants of former cataclysmically disrupted satellites, or ejecta/dust from outer saturnian satellites like Phoebe, or interplanetary particles—has been uncertain. A possible internal (endogenic) origin has also been suggested (*30*) on the basis of dark markings beyond the contiguous dark area. Additionally, an origin by eruption of dark material, followed by deposition of brighter ice, followed by impact erosion has been proposed (*31*). Theories of formation by impacts (*26*) predict that albedo markings in craters and other topography near the edges of the dark area should be asymmetrically placed on topography because of grazing angles of impacts away from the apex of motion.

Cassini images show that Cassini Regio is heavily cratered, with at least three large impact structures with diameters >350 km. The region within ~20° latitude is, at the resolution of Cassini images, entirely darkened (except for slightly brighter crater rims), yet shows no obvious signs of volcanic flooding (Figs. 6 and 7). The largest basin is near 15°N, 30°W in the northeastern part of Cassini Regio (Fig. 6, A and E) and has a broad inner (peak) ring ~160 km in diameter. A 400-km basin is nearby, at 35°N, 80°W (Fig. 6, A and E), and a 380-km basin is on the anti-Saturn hemisphere near 15°S, 120°W. The latter structure has a somewhat polygonal outline and may have a central rise. A fourth large circular structure, ~500 km across, is detected within the bright terrain at ~45°S, 240°W in low-resolution (50 km pixel^{-1}) and high-phase (>110°) images. It, too, exhibits a central rise. These large craters have sizes, relative to Iapetus's size, consistent with the largest craters on other icy satellites. Smaller craters, tens of kilometers in diameter, show central peaks (Figs. 6A and 7B).

Crater counts have been carried out in two sets of images in the bright terrain at high southern latitudes. For crater diameters between 70 and 90 km, the crater frequencies measured in the Cassini ISS data are within a factor of ~1.5 of that found from earlier counts performed on Voyager data at a different location (*32*). This crater density confirms the old age of the bright part of the surface. Counts within the dark region are in progress. It is clear from visual inspection that it is also heavily cratered and thus old (Fig. 6A and Fig. 7, B to D). Whatever darkened the leading hemisphere of Iapetus postdated the formation of the craters within Cassini Regio.

Cassini images show substantial limb topography, relative to a reference sphere, that confirms earlier measurements from low-resolution Voyager images of unusually high topography for an object of this size (*33*). Cassini images from October 2004 captured several aligned peaks that had been identified in Voyager 2 images of the limb of Iapetus, on the anti-Saturn side, between ~185°W and ~215°W and within the dark terrain (Fig. 6, B to D). December 2004 images reveal that these aligned peaks are part of a near-equatorial ridge system that extends for more than 110° in longitude (Fig. 6, A and E, and Fig. 7, C and D). Parts of this ridge system rise more than 20 km above the surrounding plains, as determined from limb measurements. This ridge system includes sections that have sets of isolated peaks (Fig. 6, D to F), ridge segments more than 200 km in length, and

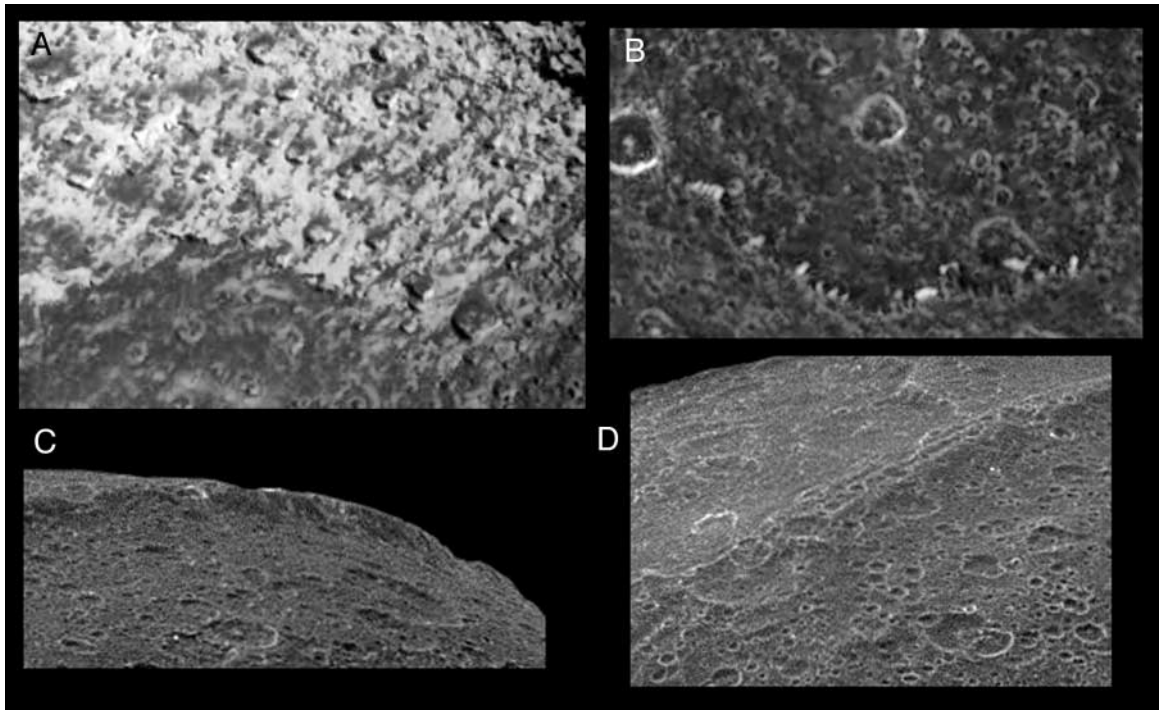


Fig. 7. Details of leading side of Iapetus. Contrast has been enhanced in all images. (**A**) Transition from Cassini Regio to polar region. Dark streaks follow trend from the center of Iapetus's leading side. Portion of image N1483173746; centered at 53°N, 74°W. North is approximately up in center of frame. (**B**) Bright streamers on north-facing slopes along rim of impact basin, bright crater rims, and some brighter central peaks. Portion of image N1483173746, centered at 25°N, 87°W, north approximately up. Bright streamers are visible on the north-facing slopes, bright crater rims, and some brighter central peaks of craters. (**C**) Equatorial ridge. Portion of N1483174305; center of ridge in this view is at 1°S, 50°W. View is ~200 km across. (**D**) Equatorial ridge. Portion of image N1483174398; centered at 3°S, 118°W. Ridge in some sections shows multiple parallel structures and many superposed impact craters.

some sections with three nearly parallel ridges. The ridge is cut by impact craters in some places, which indicates that it is not the most youthful surface feature (Fig. 7, C and D). The equatorial position of this feature suggests an origin related to the overall shape or changes in spin state of Iapetus. Iapetus is currently tidally locked into a synchronous 79-day spin. Internal energy losses will drive the bulge represented by the ridge, if not isostatically compensated, to lie along the equator, even if formed elsewhere relative to the body's spin vector (*33*). Despinning can cause complex tectonic patterns (*34*). However, the Iapetus ridge shows no characteristics or patterns comparable to the tectonic patterns seen in despinning models.

Some craters within the northern and southern transitional terrains between ~ 20° and 60° latitude show bright poleward-facing and dark equator-facing walls that are incompatible with the contrasts expected from the illumination conditions. This albedo polarity was noted in Voyager data (*29*), but Cassini images provide a higher resolution view of the albedo variations on those crater walls that exhibit this effect (Figs. 6A and 7). They show bright rims and central peaks of some craters within Cassini Regio and, in some instances, discrete downslope bright streamers occurring preferentially on north-facing slopes. An albedo asymmetry is continued into the brighter areas, where albedo variations are manifested chiefly as dark streaks on the brighter material, some associated with specific craters. The albedo contrasts make it difficult to discern the topographic outlines of many of the craters within the transition zone between ~40°N and 60°N. The orientations of the streaks away from the center of the leading side of Iapetus (which points in the direction of the apex of its orbital motion around Saturn) suggests deposition by some long-distance, ballistic transport of material (of either endogenic or exogenic origin), which either directly coated the surface of Iapetus and/or perhaps eroded or otherwise altered the pre-existing surface. The bright downslope markings on steeper crater walls, and the correlation of high albedo with regions of steeper slopes (e.g., crater rims and central peaks) within Cassini Regio indicate that the dark material is not thick (>1 km) deposits. These characteristics are consistent with some of the models of exogenic origin of Cassini Regio (*26*).

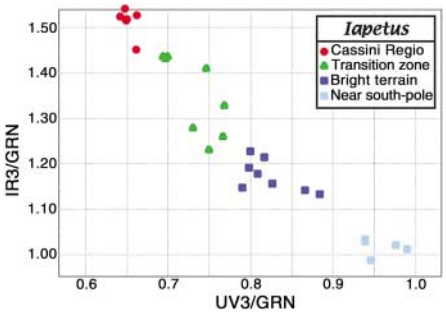


Fig. 8. Color-ratio diagram of locations on Iapetus taken at 54° phase angle; data from images N146837052, N1468370680, and N1468370800; spot sizes vary from 3 to 10 pixels across. Central wavelengths of the UV3, GRN, and IR3 color filters are 338, 568, and 930 nm, respectively. Data in the dark terrain are from 60°W to 140°W; high-latitude data are from 20°W to 150°W. The data's linear trend suggests mixing of two end members: Cassini Regio with a red spectrum and the south polar region with a flat spectrum.

Nonethess, the presence of a major geological feature of internal origin—the equatorial ridge—bisecting the leading hemisphere might suggest a genetic relationship between the ridge and the darkening of Cassini Regio. Such a relationship would require, to begin with, an internal heat source sufficient to drive eruptive volcanism extensive enough to cover an entire hemisphere. Although some degree of volcanism early in Iapetus's history might have been possible, there is no presently known heat source (such as conventional tidal heating) large enough for such extensive resurfacing of an entire hemisphere postdating the cratering epoch. Until a mechanism is found to generate substantial volcanic activity on Iapetus after the formation of its cratered surface, the Cassini evidence weighs in favor of an exogenic origin for Cassini Regio.

Data obtained through wide-band filters with central wavelengths between 380 nm and 930 nm (Fig. 8) show that there is a continuous range of colors between the darkest areas and the bright poles. Cassini Regio's color is uniformly red over 60° of longitude. Going south from the transition zone, the colors show a continuous progression of values to a nearly flat spectrum near the pole (Fig. 8). This behavior is consistent with spatial mixtures of two material end members, a dark non-ice and a bright water-ice component, at scales below the image resolution.

In summary, Phoebe is a body of rock and ice that exhibits an unexpected complexity of surface geology due to gravity-driven processes operating on a heavily cratered and compositionally heterogeneous surface. Iapetus's dark region is ancient, is bisected by a large equatorial ridge system, and shows evidence of ballistic emplacement of materials across its leading hemisphere.

References and Notes

1. J. A. Burns, in *Satellites*, J. A. Burns, M. S. Mathews, Eds. (Univ. Arizona Press, Tucson, 1986), pp. 117–158.
2. D. P. Simonelli *et al.*, *Icarus* **138**, 249 (1999).
3. T. C. Owen *et al.*, *Icarus* **149**, 160 (2001).
4. B. J. Buratti *et al.*, *Icarus* **155**, 375 (2002).
5. C. C. Porco *et al.*, *Space Sci. Rev.* **115**, 363 (2004).
6. J. M. Bauer, B. J. Buratti, D. P. Simonelli, W. M. Owen Jr., *Astrophys. J.* **610**, L57 (2004).
7. T. R. Colvin, M. E. Davies, P. G. Rogers, J. Heller, NASA STI/Recon. Tech. Rep. N-2934-NASA (1989).
8. Views of the shape models and other supporting materials are available on *Science* Online.
9. R. A. Jacobson *et al.*, *Bull. Am. Ast. Soc.* **35**, Abs. 15.02 (2004).
10. T. C. Owen *et al.*, *Icarus* **139**, 379 (1999).
11. M. E. Brown, *Astron. J.* **119**, 977 (2000).
12. R. N. Clark *et al.*, *Nature*, in press (2005).
13. G. Schubert *et al.*, in *Satellites*, J. A. Burns, M. S. Mathews, Eds. (Univ. Arizona Press, Tucson, 1986), pp. 224–292.
14. W. B. McKinnon, J. I. Lunine, D. Banfield, in *Neptune and Triton*, D. P. Cruikshank, Ed. (Univ. Arizona Press, Tucson, 1996), pp. 807–877.
15. G. Neukum, H. Dietzel, *Sci. Lett.* **12**, 59 (1971).
16. G. Neukum, B. A. Ivanov, in *Hazards Due to Comets and Asteroids*, T. Gehrels, Ed. (Univ. Arizona Press, Tucson, 1994), pp. 359–416.
17. D. Nesvorny, J. L. A. Alvarellos, L. Dones, H. F. Levison, *Astron. J.* **126**, 398 (2003).
18. B. Gladman *et al.*, *Nature* **412**, 163 (2001).
19. G. Neukum *et al.*, *Lunar Planet. Sci. Conf.* **29**, 1742 (1998).
20. J. M. Hahn, R. Malhotra, *Astron. J.* **117**, 3041 (1999).
21. H. J. Melosh, *Impact Cratering: A Geologic Process* (Oxford Univ. Press, New York, 1989).
22. P. C. Lee *et al.*, *Icarus* **120**, 87 (1996).
23. The Cassini Regio was named after the discoverer of Iapetus, Jean-Dominique Cassini.
24. D. Morrison *et al.*, *Icarus* **24**, 17 (1975).
25. D. P. Cruikshank *et al.*, *Icarus* **53**, 90 (1983).
26. J. F. Bell, D. P. Cruikshank, M. J. Gaffey, *Icarus* **61**, 192 (1985).
27. B. J. Buratti, J. Mosher, *Icarus* **115**, 219 (1995).
28. R. G. Tabak, W. M. Young, *Earth Moon Planets* **44**, 251 (1989).
29. T. Denk, G. Neukum, *Lunar Planet. Sci. Conf.* **31**, 1660 (2000).
30. B. A. Smith *et al.*, *Science* **215**, 504 (1982).
31. P. D. Wilson, C. Sagan, *Icarus* **122**, 92 (1996).
32. T. Denk *et al.*, *Lunar Planet. Sci. Conf.* **31**, 1596 (2000).
33. S. J. Peale, in *Planetary Satellites*, J. A. Burns, Ed. (Univ. Arizona Press, Tucson, 1977), pp. 87–112.
34. J. B. Pechman, H. J. Melosh, *Icarus* **38**, 243 (1979).
35. G. Neukum, B. A. Ivanov, W. K. Hartmann, in *Chronology and Evolution of Mars*, W. K. Hartmann, J. Geiss, R. Kallenbach, Eds. (Kluwer, Dordrecht, Netherlands, 2001), pp. 53–86.
36. G. Neukum, in *The Three Galileos: The Man the Spacecraft, the Telescope*, C. Barbieri, J. H. Rahe, T. V. Johnson, Eds. (Kluwer, Dordrecht, Netherlands, 1997), pp. 201–212.
37. We acknowledge the individuals at CICLOPS and across the Imaging Team who helped in planning and designing the Phoebe and Iapetus sequences and in keeping the imaging operations running smoothly. This paper is dedicated to the memory of Damon Simonelli, who developed much of the pre-Cassini understanding of Phoebe's surface. This work has been funded by NASA/Jet Propulsion Laboratory, the UK Particle Physics and Astronomy Research Council, the German Aerospace Center (DLR), and Université Paris VII Denis Diderot, Commissariat à l'Energie Atomique, Astrophysique Interactions Multiechelles, France.

Supporting Online Material

www.sciencemag.org/cgi/content/full/307/5713/1237/DC1
Figs. S1 and S2

29 November 2004; accepted 12 January 2005
10.1126/science.1107981

Cassini Imaging Science: Initial Results on Saturn's Atmosphere

C. C. Porco,[1]* E. Baker,[1] J. Barbara,[2] K. Beurle,[3] A. Brahic,[4] J. A. Burns,[5] S. Charnoz,[4] N. Cooper,[3] D. D. Dawson,[6] A. D. Del Genio,[2] T. Denk,[7] L. Dones,[8] U. Dyudina,[9] M. W. Evans,[3] B. Giese,[10] K. Grazier,[11] P. Helfenstein,[5] A. P. Ingersoll,[9] R. A. Jacobson,[11] T. V. Johnson,[11] A. McEwen,[6] C. D. Murray,[3] G. Neukum,[7] W. M. Owen,[11] J. Perry,[6] T. Roatsch,[10] J. Spitale,[1] S. Squyres,[5] P. Thomas,[5] M. Tiscareno,[5] E. Turtle,[6] A. R. Vasavada,[11] J. Veverka,[5] R. Wagner,[10] R. West[11]

The Cassini Imaging Science Subsystem (ISS) began observing Saturn in early February 2004. From analysis of cloud motions through early October 2004, we report vertical wind shear in Saturn's equatorial jet and a maximum wind speed of ~375 meters per second, a value that differs from both Hubble Space Telescope and Voyager values. We also report a particularly active narrow southern mid-latitude region in which dark ovals are observed both to merge with each other and to arise from the eruptions of large, bright storms. Bright storm eruptions are correlated with Saturn's electrostatic discharges, which are thought to originate from lightning.

During the Voyager flybys of the giant planets in the 1980s, Saturn was probably the windiest planet in the solar system (*1*). Its equatorial jet flowed eastward at 470 m s^{-1} (*2*), which is more than 10 times the speed of Earth's tropospheric jet streams. The jets are powered partly by solar energy and partly by internal heat; their sum is ~1.8 times the globally averaged absorbed solar energy. The solar irradiance is 1% that at Earth. Saturn's atmosphere is mostly H_2 and He, with small amounts of CH_4, NH_3, PH_3, and presumably H_2S and H_2O, although the last two are frozen out below the cloud tops and are undetectable by the usual remote sensing methods. On the basis of solar elemental abundances, the inferred abundance of oxygen (and hence of water) in the deep atmosphere is sufficient to support moist convection, precipitation, and lightning.

On Saturn, the wind is defined as the velocity relative to the System III reference frame (*3*), which rotates at a fixed period of 10 hours 39 min 24 s. This was the period of Saturn's kilometric radio (SKR) emissions at the time of the Voyager flybys in 1980–1981. Since then the SKR period has varied by 1% (*4*) and is currently 10 hours 45 min 45 s (*5*). In addition, Hubble Space Telescope (HST) observations from 1996 to 2004 indicate that the equatorial jet slowed to ~275 m s^{-1} relative to System III (*2*, *6*, *7*). Any real slowing of the winds might be connected to the extreme seasonal cycles of insolation (sunlight incident on a horizontal surface) as the shadow of the rings moves from one hemisphere to the other (*8*, *9*), or to the extreme outbursts of cloud activity, the last of which disrupted the appearance of the equatorial zone (EZ) between 1990 and 1997 (*10*–*12*). It is also possible that the winds are constant in time but the altitude of the visible clouds has increased, revealing slower speeds at higher levels (*10*, *12*, *13*). We focus on the last hypothesis.

The ISS cameras imaged Saturn's atmosphere regularly between 6 February and 12 May 2004, and again on 29–30 May 2004 during Cassini's final approach to Saturn (*14*) and in September during the first orbit of Saturn. The long time baseline and the wide spectral range of the ISS cameras, from the near-infrared (near-IR) to the ultraviolet, are ideal for discriminating clouds and hazes at various altitudes and for monitoring winds, vortices, and evolving cloud structures (*15*, *16*). Coverage in the southern hemisphere was complete, but the northern hemisphere was obscured by the shadow of Saturn's rings (*17*).

We measured wind speeds by tracking the displacements of spots and cloud/haze features

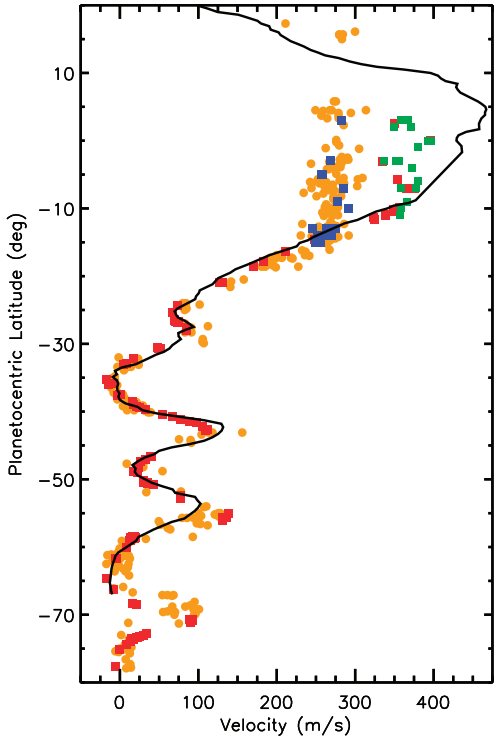


Fig. 1. Zonal winds (positive eastward) in Saturn's atmosphere. Winds in the equatorial region vary over time and with image wavelength. Measurements made by Voyagers 1 and 2 in 1980–1981 through the green filter (*2*) are the fastest (solid line). Winds derived from 1996–2004 HST images (*7*) are slower and independent of wavelength (orange circles). Cassini measurements at 750 nm, made by two authors (red and green squares), fall between those of Voyager and HST. Cassini measurements at 727 nm (blue squares) are close to those of HST. Winds south of 15°S show few changes from 1980–1981 (Voyager, solid line) (*2*) to 1996–2004 (HST, orange circles) (*7*) to 2004 (Cassini at 750 nm, red squares). Voyager observations extend only to 65°S and do not reveal the jet near 70°S.

[1]Cassini Imaging Central Laboratory for Operations, Space Science Institute, 4750 Walnut Street, Suite 205, Boulder, CO 80301, USA. [2]Goddard Institute for Space Studies, NASA, 2880 Broadway, New York, NY 10025, USA. [3]Astronomy Unit, Queen Mary, University of London, London E1 4NS, UK. [4]Centre d'Etudes de Saclay, Université Paris 7, L'Orme des Merisiers, 91191 Gif-sur-Yvette Cedex, France. [5]Department of Astronomy, Cornell University, Space Sciences Building, Ithaca, NY 14853, USA. [6]Department of Planetary Sciences, University of Arizona, 1629 East University Boulevard, Tucson, AZ 85721, USA. [7]Institut für Geologische Wissenschaften, Freie Universität, 12249 Berlin, Germany. [8]Department of Space Sciences, Southwest Research Institute, 1050 Walnut Street, Suite 400, Boulder, CO 80302, USA. [9]Division of Geological and Planetary Sciences, California Institute of Technology, 150-21, Pasadena, CA 91125, USA. [10]Institute of Planetary Research, German Aerospace Center, Rutherfordstrasse 2, 12489 Berlin, Germany. [11]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA.

*To whom correspondence should be addressed. E-mail: carolyn@ciclops.org

in Cassini images separated by either 10.5 or 21 hours (*18*). The speeds of the zonal winds poleward of 15°S latitude are similar to those measured during the Voyager and HST eras (Fig. 1). The zonal (eastward) velocity has local maxima at 27°, 43°, 55°, and 70°S. There are zonal velocity minima at 35°, 48°, and 65°S. Images taken in the deep-sounding continuum CB2 (750 nm) filter give equatorial jet speeds of 325 to 400 m s^{-1}, which falls between the Voyager and HST estimates. Images taken in the gaseous methane absorption band MT2 (727 nm) filter, which senses higher altitudes, give equatorial jet speeds of 250 to 300 m s^{-1}, in agreement with the HST result (Fig. 1). This relation between the measured wind speed and the altitude to which the filter is sensitive was first noticed in HST images after the 1990 equatorial disturbance and was interpreted as evidence of vertical shear in the zonal winds (*10*). For the HST data from 1996 to 2004, the large high-contrast features in the EZ emerged in the strong methane band (890 nm). The authors of the 1996–2004 study estimate that the EZ features' average level was $P \sim 45$ mbar in 2003 and was ~300 mbar during the Voyager era (*7*).

The equatorial velocities from the Voyager, HST, and Cassini data imply vertical wind shear. The MT2 filter is centered on a methane absorption band, so clouds are visible only if they are at high altitude—above most of the methane gas. The CB2 filter is outside the methane bands, so clouds are visible at all levels. Optically thin clouds at high altitude will be more visible in the MT2 filter, where the methane gas provides a dark background, and they will be less visible in the CB2 filter, where bright deep clouds provide the background. Figure 2 shows an example of different features appearing in different filters at the same time and place on the planet.

Wind speed observations from Voyager were determined by tracking clouds seen in

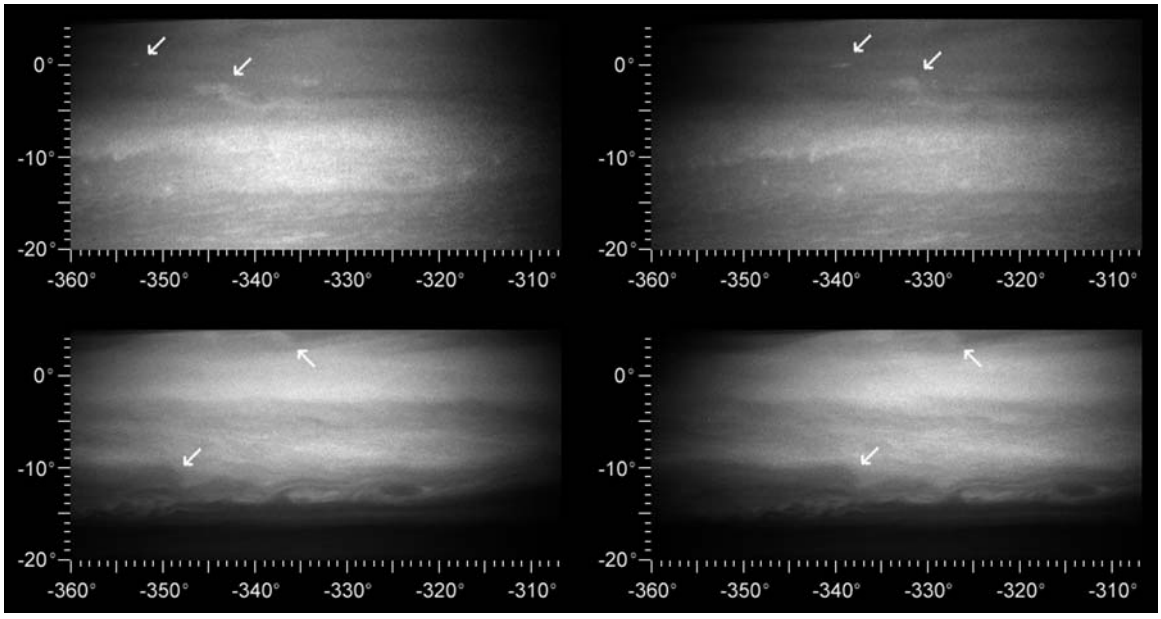


Fig. 2. Near-simultaneous views in different filters of the same place on the planet. The upper left and right images use the CB2 filter (750 nm), where the gases absorb weakly; the lower left and right images use the MT2 filter (727 nm), where methane absorbs more strongly. The left images were taken within 1 min of each other and cover the same latitude-longitude range. They were taken 10 hours 30 min before the right images, which were also taken within 1 min of each other and also cover the same latitude-longitude range. The velocities measured by tracking the features (arrows) in the CB2 images (top) are ~100 m s^{-1} faster than those in the MT2 images (bottom).

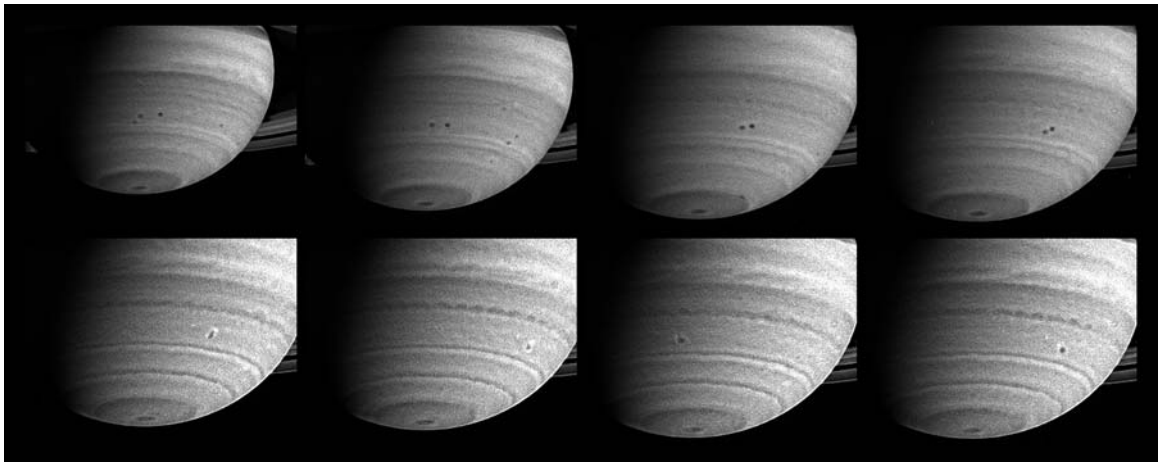


Fig. 3. Two storms, with diameters close to 1000 km, were captured in the act of merging during Cassini's approach to Saturn. Both storms, which appear as spots in the southern hemisphere, were seen moving westward, relative to System III, for about a month before they merged on 19–20 March 2004. The series of eight images shown here was taken between 22 February and 22 March 2004; the image scale ranges from 381 km (237 miles) to 300 km (186 miles) per pixel. All images have been processed to enhance visibility. The top four frames, spanning 26 days, are portions of narrow-angle camera images that were taken through a filter accepting light in the near-IR region of the spectrum centered at 619 nm, and show two spots approaching each other. Both storms are within half a degree of 36°S latitude and sit in an anticyclonic shear zone, which means that the flow to the north is westward relative to the flow to the south. Consequently, the northern storm moves westward at 11 m s^{-1}, a slightly greater rate than the southern one (6 m s^{-1}). The storms drift with these currents and engage in a counterclockwise dance before merging with each other. The bottom four frames are from images taken on 19, 20, 21, and 22 March, respectively, in a region of the spectrum visible to the human eye and illustrate the storms' evolution. Just after the merger, on 19 March, the new feature is elongated in the north-south direction, with bright clouds on either end. Two days later, on 22 March, it has settled into a more circular shape and the bright clouds have spread around the circumference to form a halo.

the Voyager green filter, which senses approximately the same altitude as the CB2 filter. The 100 m s^{-1} difference in equatorial speeds between the CB2 and MT2 filters is consistent with winds decreasing with height coupled with thin high clouds. Retrievals of haze altitudes support the idea that the 1990–1997 equatorial storms put haze particles at higher altitudes than during the Voyager years. The pressure at the top of the main haze layer appears to have decreased from ~200 mbar in 1979 (*19*) to about 70 mbar in 2002 (*20*).

Retrievals of temperature support the idea that the zonal wind speed decreases with height. The Cassini Composite Infrared Spectrometer (CIRS) team (*13*) observed an increase in temperature toward the south pole, which implies that the zonal winds decay with height according to the thermal wind equation. This equation is singular at the equator, so the wind shear profile stops at 5°S. If the Voyager wind profile holds at the 500-mbar level, at 5°S the winds will reach 275 m s^{-1} at the 3-mbar level, and at 10°S they will reach 275 m s^{-1} at the 30-mbar level. These altitudes are higher than the 45- and 70-mbar estimates given earlier for the stratospheric haze, but they are within the combined uncertainty of estimating the altitude of features from methane band filters and applying the thermal wind equation within 5° of the equator.

The westward jet at 35°S planetocentric latitude (Fig. 1) was the most active region of the planet during Cassini's long approach to Saturn in the first half of 2004 (*7*). Voyager 2 observed similar activity in 1980–1981 near the westward jet at 34°N (*21*). "Activity" is here defined to mean having a dozen or more compact oval spots with diameters larger than 500 km and lifetimes longer than a week. At 35°S, most of the spots sat in the shear zone just to the south of the jet. Spots in the jet moved westward at speeds up to 25 m s^{-1}. Spots a few degrees to the south moved eastward at comparable speeds. The spots themselves rotated counterclockwise (anticyclonic in the southern hemisphere), like ball bearings between conveyor belts, with the same sense of rotation as the ambient shear flow. It was not possible to determine the sense of rotation for all the spots, especially the smallest ones, but the impression from movies of time-lapse images (movies S1 and S2) is that they were all anticyclonic. The jet at the zonal velocity minimum at 48°S was also active and exhibited the same phenomena, with anticyclonic vortices rolling in the anticyclonic shear zone to the south of the jet.

The interaction of westward-moving and eastward-moving spots in adjacent latitude bands sometimes produced a merger (Fig. 3) or a near-merger, with bright material suddenly forming between the two spots or around them. Further analysis of the images taken in different spectral filters will show whether the bright material consists of particles of a different composition, size, or altitude. The compact spots were mostly dark, but they often had bright halos. The new spot that resulted from a merger sometimes continued on for weeks and sometimes split apart after a few days. Because mergers represent a net decrease in the number of spots, there must be another process that causes the spots to form. Figures 4 and 5 show one possibility. In Fig. 4, the first six frames show a large eruption of bright white clouds in the 35°S latitude band. The eruption, which suggests moist convection, later produces three dark stable spots where there were none before. In Fig. 5, days 272 and 274 record the curling of a dark filament on its way to becoming a dark spot. These observations may indicate that moist convection, through the injection of vorticity at cloud level, supplies energy to maintain the jets on jovian planets (*22*).

Linear bands encircle the planet, each with its own mottled texture. The edges of the bands are wave-like and often break into turbulent eddies. On the largest scale, rotational symmetry is the dominant feature (Fig. 3). A curious manifestation of this symmetry occurs at the south pole, where a small (2000 km diameter) feature sits undisturbed. The precise alignment suggests that the perturbing forces are small.

One storm at 35°S produced eruptions of white material in mid-July and again in mid-September (Fig. 5). During these eruptions, the Radio and Plasma Wave Science (RPWS) instrument (*3*, *5*, *23*) was detecting Saturn electrostatic discharges (SEDs), which are short bursts—tens of milliseconds in duration—with large bandwidth, spanning 2 to 40 MHz in frequency. The SEDs are different from SKR radiation, although both are radio waves. The SED bursts are grouped into episodes that recur with a distinct periodicity (~10 hours 40 min), which is approximately equal to the System III rotation period. Voyagers 1 and 2 also detected SEDs, which were also grouped into episodes, but the recurrence period was ~10 hours 10 min (*3*). The bursts may have originated in the rings, and they may have originated in the atmosphere (*24*). The Cassini data favor an atmospheric origin (i.e., lightning).

The Voyager and Cassini periods for the SEDs are close to the minimum and maximum recurrence intervals, respectively, for atmospheric features crossing the center of the planetary disk—the central meridian (CM)—as viewed by a stationary observer. The shorter period is characteristic of the high-speed equatorial jet. The longer period is characteristic of the spots just south of the westward jet at 35°S, where the storm erupted in July and September 2004. We matched the phase of the SED bursts with the times that the storm crossed the CM over a period spanning 58 rotations of the planet, which is more than 25 days. A constant relative phase

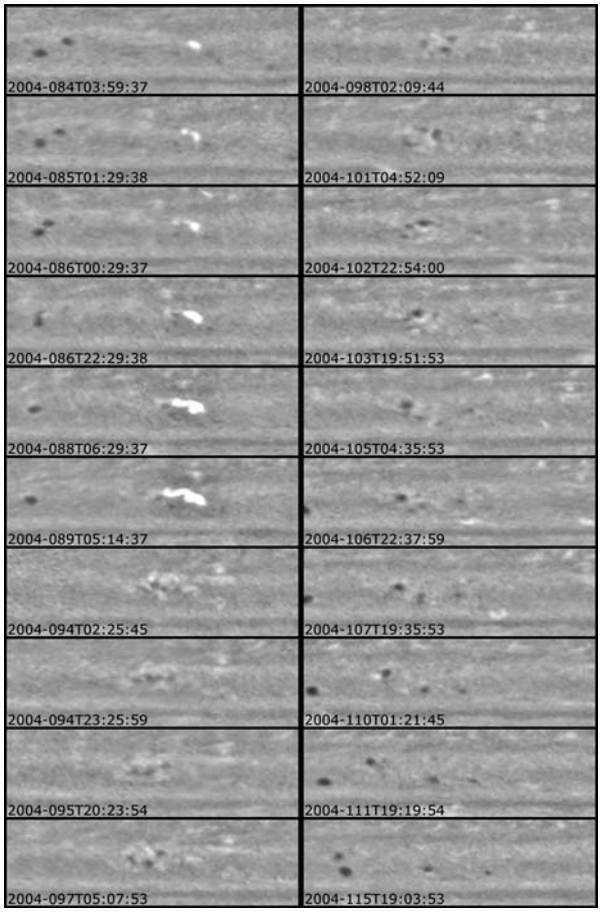


Fig. 4. Spots merging and spots arising. Each panel shows a region in the vicinity of the westward jet at 35°S taken from one of the maps that constitute the cylindrical movie (movie S2) of Saturn's southern hemisphere. The time shown is that of the first of the six images that make up each cylindrical map (*14*); the actual time the region was imaged is within one rotation period after the time shown. In the first six frames, two dark spots merge (on approximately day 86, 26 March 2004) after spiraling toward each other in a counterclockwise direction. In the last seven frames, three dark spots arise from the remnants of the bright white storm seen in the earlier frames and spin around each other. Two of the newly developed spots merge between day 107 and day 110.

would suggest that the radio bursts are originating in the atmosphere and are probably caused by lightning discharges associated with the storm.

The results of this analysis are shown in Fig. 6. The SED "on" and "off" periods each last for ~5 hours, or ~0.5 planetary rotations. The bursts begin ~0.4 planetary rotations before the storm's CM crossing and end ~0.1 rotations after it (Fig. 6, insets). The phase of the SED episodes relative to the CM crossings remains constant and is the same for both outbursts, indicating that they are related. The maxima in the amplitude of the SED episodes (Fig. 6) at days 256 and 270 are correlated with an unusual brightening of the storm (Fig. 5), although the dynamic range of SED variation is greater than that of the storm brightness.

The peak SED intensity occurs ~0.2 planetary rotations (~72°) before the CM crossing, which suggests that the lightning source is ~72° to the east of the visible cloud, which trails off to the west because of wind shear. For this explanation to work, the lightning source would have to be deeper than the visible clouds and the wind would have to be more westward at higher altitude. Then the SEDs would lead the visible storm, as observed. Voyager 2 saw similar evidence of vertical shear in the zonal wind at almost the same latitude, but there the inference was that the winds were more eastward at higher altitude (*1*).

The two Voyagers detected low-frequency SED bursts—below 5 MHz—only when the spacecraft was above the night side of the planet (*3*, *25*). If the ionospheric plasma density near the noon meridian was $\geq 3 \times 10^5$ cm^{-3}, escape of atmospheric radio bursts below 5 MHz would be prevented (*3*). If the nightside ionosphere had low-density regions or holes, the observed low-frequency emission could escape. This could explain the phase shift observed by Cassini, because the storm was in darkness before it crossed the CM and was in sunlight after CM crossing. The ionospheric cutoff would explain why SEDs appeared to turn off after the storm crossed the CM, but it does not explain why the SEDs began while the storm was still below the horizon (i.e., more than 0.25 rotations before CM crossing).

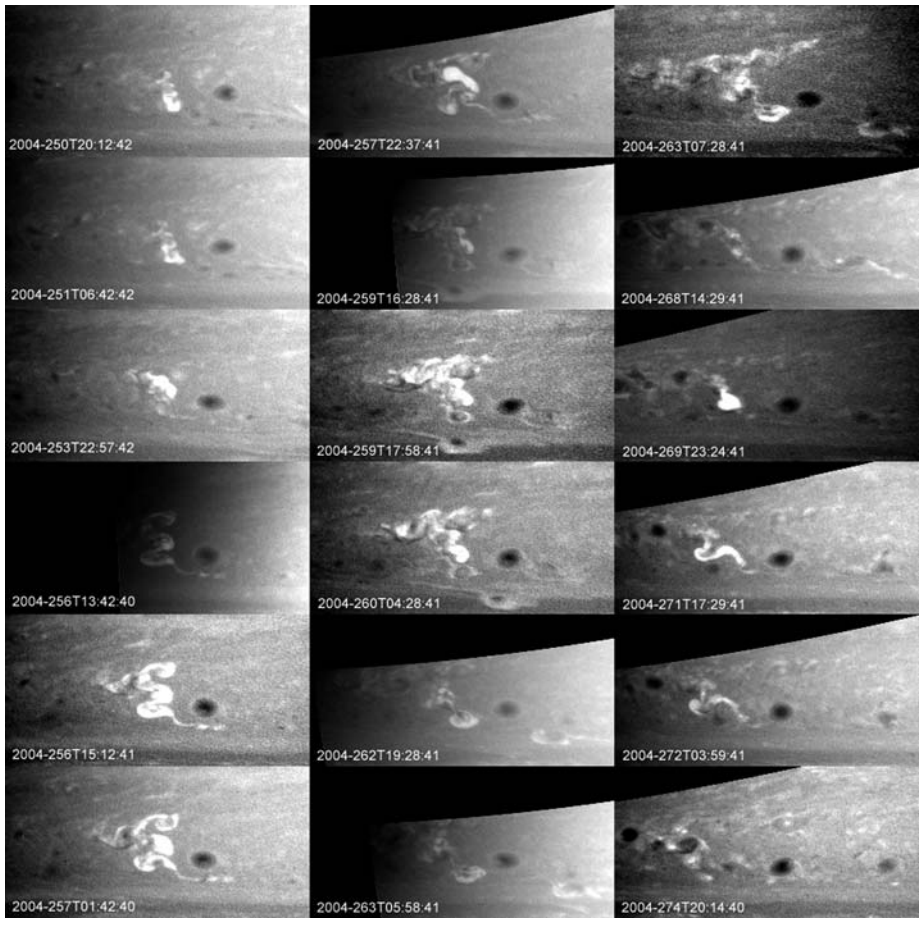


Fig. 5. Time sequence of the September storm at 35°S latitude. The sequence begins on day 250 (6 September 2004) and ends on day 274. The storm reaches its peak intensity in visible light on days 255 to 260. After fading for a few days, it returns as a high-contrast bright spot on day 269. Between day 272 and day 274, a dark filament appears and is observed curling up on its way to becoming a new dark spot.

References and Notes

1. A. P. Ingersoll, R. F. Beebe, B. J. Conrath, G. E. Hunt, in *Saturn*, T. Gehrels, M. S. Matthews, Eds. (Univ. of Arizona Press, Tucson, AZ, 1984), pp. 195–238.
2. A. Sánchez-Lavega, J. F. Rojas, P. V. Sada, *Icarus* **147**, 405 (2000).
3. M. L. Kaiser *et al.*, in *Saturn*, T. Gehrels, M. S. Matthews, Eds. (Univ. of Arizona Press, Tucson, AZ, 1984), pp. 378–415.
4. P. H. M. Galopeau, A. Lecacheux, *J. Geophys. Res.* **105**, 13089 (2000).
5. D. A. Gurnett *et al.*, *Science* **307**, 1255 (2004); published online 16 December 2004 (10.1126/science.1105356).
6. A. Sánchez-Lavega, S. Pérez-Hoyos, J. F. Rojas, R. Hueso, R. G. French, *Nature* **423**, 623 (2003).
7. A. Sánchez-Lavega, R. Hueso, S. Pérez-Hoyos, J. F. Rojas, R. G. French, *Icarus* **170**, 519 (2004).
8. The axial tilts of Jupiter and Saturn are 3.13° and 26.73°, respectively. Saturn's orbital period (and seasonal cycle) is 29.4 years.
9. B. Bézard, D. Gautier, B. Conrath, *Icarus* **60**, 274 (1984).
10. C. D. Barnet, J. A. Westphal, R. F. Beebe, L. F. Huber, *Icarus* **100**, 499 (1992).
11. A. Sánchez-Lavega *et al.*, *Nature* **353**, 397 (2000).
12. A. Sánchez-Lavega *et al.*, *Science* **271**, 631 (1996).
13. F. M. Flasar *et al.*, *Science* **307**, 1247 (2004); published online 23 December 2004 (10.1126/science.1105806).
14. A global map of Saturn, obtained by imaging Saturn's disk every 60° longitude as the planet rotated below the spacecraft, was acquired roughly every other rotation of Saturn. The spatial resolution of the narrow-angle camera ranged from 425 to 156 km per pixel. Images were navigated on the planetary limb, calibrated radiometrically, assembled into spatial mosaics, and mapped with a simple cylindrical projection. Illumination effects were removed by dividing the brightness value by the cosine of the solar incidence angle. Images from 29–30 May were part of a campaign to view Saturn at near-IR wavelengths.
15. C. C. Porco *et al.*, *Space Sci. Rev.* **115**, 363 (2004).
16. C. C. Porco *et al.*, *Science* **299**, 1541 (2003).
17. The solar phase angle was ~65°. The subsolar and subspacecraft planetocentric latitudes were –25° and –16°, respectively.
18. Winds were measured on images from 10–12 May, 29–30 May, and 6–7 September 2004. Spatial resolutions were ~157 km per pixel, ~104 km per pixel, and ~53 km per pixel, respectively, resulting in maximum wind speed precisions of 2.1 m s^{-1}, 1.4 m s^{-1}, and 0.7 m s^{-1}. Limited illumination and contrast within Saturn's cloud bands made identifying features more challenging than at Jupiter. Two authors (A.R.V., J.B.) tracked features with the use of different methods. One selected the initial and final tie points by eye. The other selected the feature by eye but used a digital correlation window to objectively determine the tie points.
19. R. A. West *et al.*, *Icarus* **51**, 51 (1982).
20. T. Temma *et al.*, *Icarus,* in press.
21. L. A. Sromovsky, H. E. Revercomb, R. J. Kraus, V. E. Suomi, *J. Geophys. Res.* **88**, 8650 (1983).
22. A. P. Ingersoll, P. J. Gierasch, D. Banfield, A. R. Vasavada, Galileo Imaging Team, *Nature* **403**, 630 (2000).
23. M. Desch and the RPWS team are acknowledged in the codiscovery of the correlation between the SEDs and Saturn convective storm activity and for allowing use of the data in figure 2 of Gurnett *et al.* (*5*).
24. J. A. Burns, M. R. Showalter, J. N. Cuzzi, R. H. Durisen, *Icarus* **54**, 280 (1983).
25. M. L. Kaiser, J. E. P. Connerney, M. D. Desch, *Nature* **303**, 50 (1983).
26. We acknowledge the many individuals across the imaging team who have assisted in the design of imaging sequences and camera commands and in other vital operational and image processing tasks, in particular N. Martin, J. Riley, E. Birath, B. Knowles, C. Clark, M. Belanger, and D. Wilson. Supported by NASA/JPL, the UK Particle Physics and Astronomy Research Council, the Deutsches Zentrum für Luft- und Raumfahrt (German Aerospace Center), and Université Paris VII Denis Diderot, Commissariat a l'Energie Atomique, Astrophysique Interactions Multiechelles, France.

Supporting Online Material

www.sciencemag.org/cgi/content/full/307/5713/1243/DC1
Movies S1 and S2

23 November 2004; accepted 6 January 2005
10.1126/science.1107691

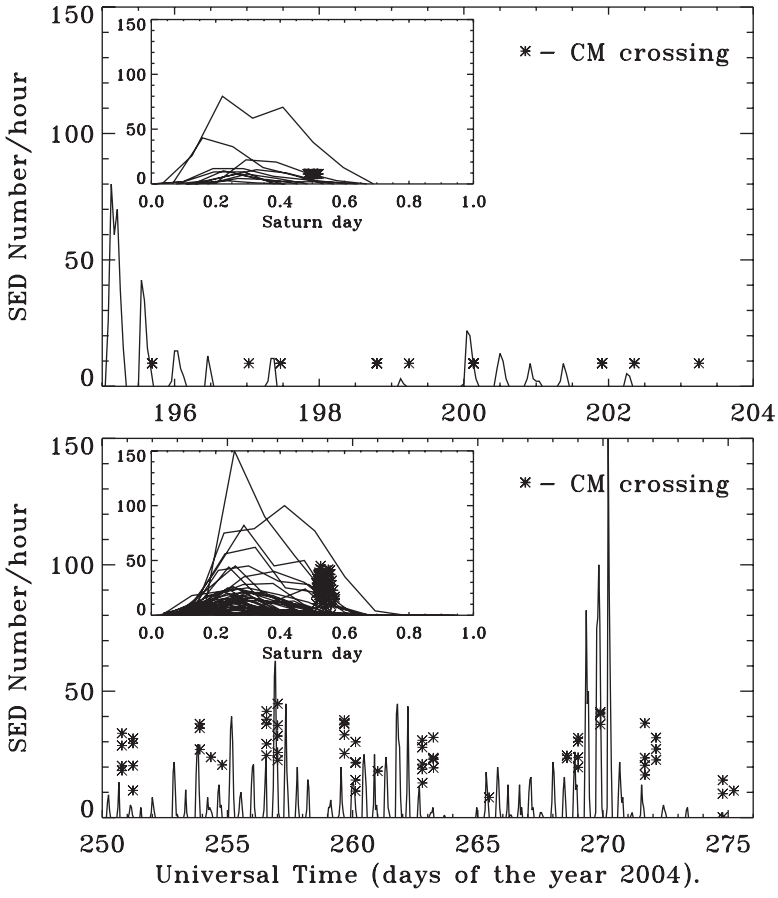


Fig. 6. Timing of the storms' CM crossings relative to the timing of the SEDs. The abscissa is Universal Time at the spacecraft; the labels refer to the start of each day. For the solid curves, the ordinate is the number of SED bursts per hour (*5*). The upper and lower panels represent different eruptions of the same storm; day 195 is 13 July 2004, and day 250 is 6 September 2004. On each image that contained the storm (a sighting), we measured the storm's longitude and computed when it would have crossed the CM on that saturnian day (Saturn rotation). Because we were looking at the dawn side of the planet, the CM crossings occurred before the storm sightings and are indicated by the positions of the asterisks along the abscissa. For images taken after day 250 (lower panel), we also computed the excess reflectivity [(storm/surroundings) – 1]. In this case the asterisk indicates not only the time of CM crossing (abscissa) but also the excess reflectivity of the storm (ordinate), in arbitrary units. Several asterisks at the same CM crossing represent several sightings of the storm on the same Saturn rotation. The different values of excess reflectivity (height of the asterisks) on the same Saturn rotation arise because the storm was observed at different positions on the half-illuminated disk. The coverage of the storms was intermittent, so gaps in the asterisks represent Saturn days when there were no observations. The coverage of SEDs was continuous, so gaps in the SED sequence represent Saturn days when there were no SEDs. The insets show the same data plotted versus Saturn time of day (Saturn's period is taken to be 10.6562 hours). Time 0.0 is when the storm is on the opposite side of the planet as viewed from Cassini; time 0.25 is the time at which the storm is rising above the horizon as seen by Cassini.

REPORT

Temperatures, Winds, and Composition in the Saturnian System

F. M. Flasar,[1]* R. K. Achterberg,[2] B. J. Conrath,[3] J. C. Pearl,[1] G. L. Bjoraker,[1] D. E. Jennings,[1] P. N. Romani,[1] A. A. Simon-Miller,[1] V. G. Kunde,[4] C. A. Nixon,[4] B. Bézard,[5] G. S. Orton,[6] L. J. Spilker,[6] J. R. Spencer,[7] P. G. J. Irwin,[8] N. A. Teanby,[8] T. C. Owen,[9] J. Brasunas,[1] M. E. Segura,[10] R. C. Carlson,[2] A. Mamoutkine,[2] P. J. Gierasch,[3] P. J. Schinder,[3] M. R. Showalter,[11] C. Ferrari,[12] A. Barucci,[5] R. Courtin,[5] A. Coustenis,[5] T. Fouchet,[5] D. Gautier,[5] E. Lellouch,[5] A. Marten,[5] R. Prangé,[5] D. F. Strobel,[13]† S. B. Calcutt,[8] P. L. Read,[8] F. W. Taylor,[8] N. Bowles,[8] R. E. Samuelson,[4] M. M. Abbas,[14] F. Raulin,[15] P. Ade,[16] S. Edgington,[6] S. Pilorz,[6] B. Wallis,[6] E. H. Wishnow[17]

Stratospheric temperatures on Saturn imply a strong decay of the equatorial winds with altitude. If the decrease in winds reported from recent Hubble Space Telescope images is not a temporal change, then the features tracked must have been at least 130 kilometers higher than in earlier studies. Saturn's south polar stratosphere is warmer than predicted from simple radiative models. The C/H ratio on Saturn is seven times solar, twice Jupiter's. Saturn's ring temperatures have radial variations down to the smallest scale resolved (100 kilometers). Diurnal surface temperature variations on Phoebe suggest a more porous regolith than on the jovian satellites.

Cassini observations of Saturn provide a detailed comparison with Jupiter, successfully studied by Galileo, which will sharpen our ideas about the formation of planetary systems. Each giant, fluid planet has its own system of orbiting moons. In addition, Saturn has a complex system of rings. All these objects are at temperatures in the range 55 to 200 K and radiate most of their energy at mid- and far-infrared wavelengths. In this part of the spectrum, most molecules have distinctive fingerprints: manifolds of lines arising from rotational or vibrational-rotational transitions. Measuring the emitted radiation of these objects therefore probes their temperatures and composition.

The composite infrared spectrometer (CIRS) on Cassini consists of two Fourier-transform spectrometers, which together measure thermal emission from 10 to 1400 cm^{-1} (wavelengths 1 mm to 7 μm) at an apodized spectral resolution selected between 0.5 and 15.5 cm^{-1} (*1*, *2*). The far-infrared interferometer (10 to 600 cm^{-1}) has a 4-mrad field

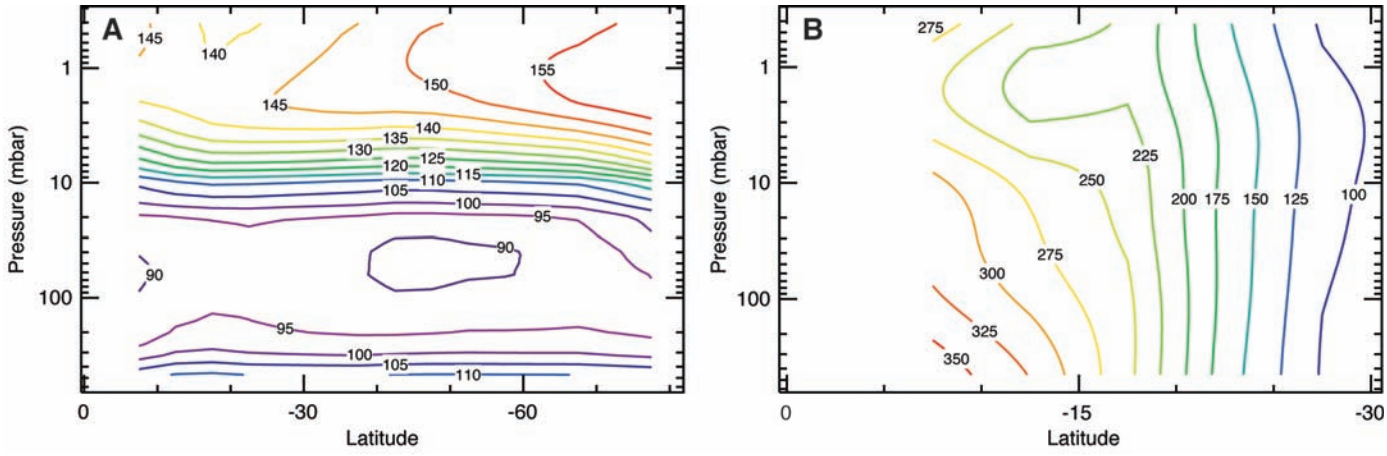


Fig. 1. Zonally averaged temperatures and zonal winds above Saturn's clouds. (**A**) Retrieved temperatures (supporting online text). The meridional cross sections have planetographic latitude (*27*) as the horizontal coordinate and pressure as the vertical coordinate (the pressure range 500 to 1 mbar corresponds to an altitude change of about 250 km). The cross sections are from data obtained in the latter half of May 2004, when the midinfrared pixel resolution was 8° latitude at 20°S and scaled as 1/cos(latitude – 20°S) at other locations. (**B**) Zonal winds, obtained from the thermal wind equation. Cloud-level winds, derived from Voyager 1 and 2 images obtained in 1980–81 (*28*), and smoothed to 8° latitude, were used as a lower boundary condition at 500 mbar. Positive numbers indicate eastward velocities. The retrieved temperature error from instrument noise is 0.1 K, and this propagates to zonal-wind errors at 5°S of 3 and 7 m s^{-1} at the 100- and 1-mbar levels, respectively; the wind error varies at other latitudes as 1/sin(latitude).

of view on the sky. The midinfrared interferometer consists of two 1 × 10 arrays of 0.3-mrad pixels, which together span 600 to 1400 cm^{-1}. The occasion of the approach of Cassini and its orbital insertion on 30 June 2004 provided the opportunity to observe Saturn, its rings, and its moons, with high spectral and good spatial resolution.

Saturn's zonal winds are enigmatic. Tracking features at the cloud-top level has indicated winds as strong as 500 m s^{-1} at the equator (*3*, *4*). These winds imply a considerable excess of axial angular momentum relative to Saturn's interior. Cloud motions observed more recently in 1996–2002, in images at visible and near-infrared wavelengths from the Hubble Space Telescope (HST), indicate that the strong equatorial winds may have slowed to ~275 m s^{-1} (*5*). The dynamic cause of such a decline is not known, but it could be related to the eruption of intense equatorial storms in 1990 and 1994 (*6*–*8*) and the zonal momentum convergence associated with propagating atmospheric waves. However, the recent HST observations (*5*) may have been biased by higher clouds near the equator. Features would be tracked higher up, where the winds are possibly slower than those tracked in the earlier studies.

During the Saturn approach phase of the Cassini spacecraft, CIRS was able to map temperatures with its midinfrared arrays over the entire southern hemisphere, which was not obscured by the rings. The spatial resolution was limited to 8° of latitude at the subspacecraft latitude (near 20° S), but this enabled the establishment of a baseline map of the southern hemisphere at moderate spatial resolution (Fig. 1). At most altitudes, the temperatures tend to increase toward the South Pole, particularly in the upper stratosphere near 1 mbar. The thermal wind equation relates the vertical shear in the zonal winds to the variation of temperatures with latitude (*9*). The horizontal resolution of temperature in our mapping is inadequate to resolve the vertical structure in the winds at mid- and high latitudes, but Fig. 1 shows a strong decay of the low-latitude zonal wind, at about 100 m s^{-1} from the cloud tops (~500 mbar) up to the 1-mbar level. If Saturn's winds had not changed from the 1980–1981 Voyager measurements, and if the recent HST observations had instead tracked higher altitude clouds, the features would need to be surprisingly high in the stratosphere. The 275-m s^{-1} winds near the equator lie at the 10-mbar level or higher (Fig. 1).

The season observed on Saturn is southern summer, about three saturnian weeks after solstice (October 2002). At 1 to 5 mbar, the South Pole is 15 K warmer than the equator. Although this might seem indicative of summer, the radiative relaxation time in Saturn's upper stratosphere is ~9.5 years (*10*), which is a significant fraction of Saturn's orbital period, 29.5 years. The atmospheric response should lag behind the seasonally modulated solar heating by $\tan^{-1}(2\pi\ 9.5/29.5) = 64°$, about a sixth of the orbital period. Simulations of the atmospheric radiative response (*11*) indicate that the South Pole should be only 5 K warmer than the equator at the time of the observations. The simulations assumed a uniform distribution of opacity, including the hydrocarbons CH_4, C_2H_2, and C_2H_6, which are the main stratospheric coolants at midinfrared wavelengths, and perhaps a nonuniform distribution of the opacity sources accounts for the discrepancy. Alternatively, there may be a dynamical cause, for example, the adiabatic heating and cooling associated with a meridianal cir-

[1]National Aeronautics and Space Administration (NASA)/Goddard Space Flight Center, Code 693, Greenbelt, MD 20771, USA. [2]Science Systems and Applications, Inc., 5900 Princess Garden Parkway, Suite 300, Lanham, MD 20706, USA. [3]Department of Astronomy, Cornell University, Ithaca, NY 14853, USA. [4]Department of Astronomy, University of Maryland, College Park, MD 20742, USA. [5]Laboratoire d'Etudes Spatiales et d'Instrumentation en Astrophysique (LESIA), CNRS–UMR 8109, Observatoire de Paris, 5 place Jules Janssen, F-91925 Meudon Cedex, France. [6]Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, CA 91109, USA. [7]Department of Space Studies, Southwest Research Institute, 1050 Walnut Street, Suite 400, Boulder, CO 80302, USA. [8]Atmospheric, Oceanic, and Planetary Physics, Clarendon Laboratory, Parks Road, University of Oxford, Oxford OX1 3PU, UK. [9]University of Hawaii, Institute of Astronomy, 2680 Woodlawn Drive, Honolulu, HI 96822, USA. [10]QSS Group, Inc., 4500 Forbes Boulevard, Suite 200, Lanham, MD 20706, USA. [11]Stanford University, c/o 245-3 NASA Ames Research Center, Moffett Field, CA 94035–1000, USA. [12]Commissariat de l'Energie Atomique, Saclay, Service d'Astrophysique, 91191 Gif-sur-Yvette Cedex, France. [13]Department of Earth and Planetary Science, Johns Hopkins University, Baltimore, MD 21218, USA. [14]NASA/Marshall Space Flight Center, SD50 National Space Science and Technology Center, Huntsville, AL 35812, USA. [15]The Laboratoire Interuniversitaire des Systèmes Atmosphériques (LISA), Université de Paris 7 and 12, CNRS–UMR 7583, 61 Avenue General de Gaulle, 94010 Créteil Cedex, France. [16]Department of Physics and Astronomy, University of Cardiff, 5 The Parade, Cardiff CF24 3YB, UK. [17]Lawrence Livermore National Laboratory and Space Sciences Laboratory, University of California, Berkeley, L-041, Livermore, CA 94551, USA.

*To whom correspondence should be addressed. E-mail: f.m.flasar@nasa.gov

†Present address: Laboratoire d'Etudes Spatiales et d'Instrumentation en Astrophysique (LESIA), CNRS–UMR 8109, Observatoire de Paris, 5 place Jules Janssen, F-91925 Meudon Cedex, France.

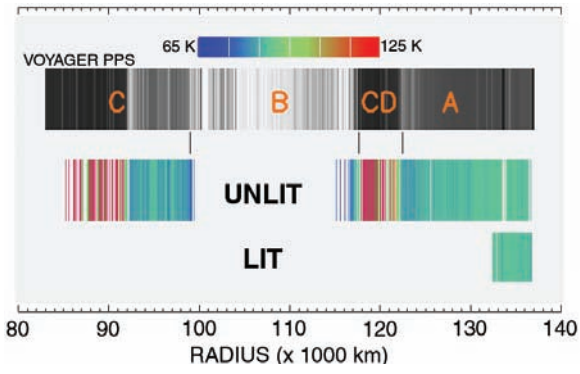


Fig. 2. Optical depth and temperature of Saturn's rings. **(Top)** Voyager photopolarimeter (PPS) radial optical depth profile from stellar occultation data taken in 1982 (*29*), spatially averaged to provide resolution comparable to the CIRS data. Light regions are optically thick, dark regions optically thin. **(Middle)** CIRS temperature profile of unlit rings. Low temperatures occur in optically thick regions, high temperatures in optically thin ones. **(Bottom)** CIRS temperature profile of lit A ring. The lit A ring temperatures are comparable to the unlit A ring temperatures in the middle strip.

culation involving upwelling at the equator and subsidence at high southern latitudes. The vertical velocities needed are small, 0.01 cm s^{-1} (supporting online text). The variation with latitude of temperatures in the tropopause region (~100 mbar) is more subdued, consistent with the radiative response when the relaxation time is long (*10*). The time-dependent response should have a smaller amplitude and should lag behind the external solar forcing by one full season. Temperatures in the upper troposphere, retrieved from Voyager infrared spectra close to the northern spring equinox, showed a warm southern hemisphere, presumably a response to the previous season's solar heating (*12*, *13*).

In the atmospheres of the giant planets, CH_4 is the principal reservoir of carbon and CH_4 photolysis is the prime source of ethane and acetylene, which are observed in CIRS spectra of Saturn (fig. S1). Because CH_4 does not condense on Saturn, it should be uniformly distributed in the upper troposphere and stratosphere. Previous determinations have yielded $[CH_4]/[H_2] = 2.0-4.5 \times 10^{-3}$ from observations in the red, near-infrared, and midinfrared bands of CH_4 (table S1). The observations in the red bands and near infrared require haze models to derive volume mixing ratios, and this adds an element of uncertainty. The midinfrared (7 μm) observation suffers from an ambiguity between CH_4 abundance and temperature and required the use of a temperature profile retrieved from Voyager 2 radio occultation soundings, which has large uncertainties at 1 mbar, in the region of the CH_4 line formation. The pure rotational lines of CH_4 in the far infrared permit a more accurate determination of CH_4, because there is redundancy in the spectrum that allows a separation of temperature and the CH_4 abundance at the line-formation level, ~400 mbar. The best fit to the spectra (fig. S2) yields a mole fraction of $4.5 \pm 0.9 \times 10^{-3}$ (1 σ) for CH_4 ($[CH_4]/[H_2] = 5.1 \pm 1.0 \times 10^{-3}$).

The new value implies that C/H on Saturn is 7 ± 2 times the solar ratio (*14*). This means that carbon is enriched on Saturn by twice the amount found on Jupiter (*15*), which is consistent with the core accretion model of giant planet formation (*16*). In this model, both Jupiter and Saturn began formation by accreting cores of about 10 to 12 Earth masses (M_E) (*17*). The subsequent collapse of surrounding solar nebula gas would bring additional heavy elements (*18*). For Jupiter, this admixture delivers ~6 M_E if one assumes that the fraction of heavy elements in a solar mixture has $Z = 0.02$ (*14*), which leads to a total mass of heavy elements of 16 to 18 M_E. For Saturn, the final heavy element mass is ~12 to 14 M_E, whereas the expected total abundance of heavy elements for a solar composition object with Saturn's mass is only 2 M_E. Thus, we can predict (*16*) the seven times enrichment that is, in fact, observed. Although the heavy element enrichments of both Jupiter and Saturn (and Uranus and Neptune as well) can be quantitatively predicted by this simple approach, the significance of the agreement must be tempered by the possibility of mere coincidence. In particular, the relative importance of material dredged from the cores and planetesimals dissolving in the envelopes of these objects remains to be worked out.

As Cassini entered orbit about Saturn, it passed above the unlit side of the main rings. During that time, CIRS obtained far-infrared spectra with spatial resolution of 100 to 200 km. After then crossing the ring plane, the lit side of the A ring was observed at 600 to 700 km resolution (table S2).

Representative temperature and optical depth spectra were derived from the data (supporting online text). As a simple model, we considered the rings to be thin, non-scattering slabs, for which the observed radiance is the product of an emissivity, a Planck function representing the ring temperature, and a filling factor described in terms of the ring optical depth. For simplicity, we assumed that the particle emissivity is unity. Temperatures derived for the unlit rings vary from 70 K to 110 K, with a strong anticorrelation between optical depth and temperature. The optically thick portions of the A, B, and C rings and of the Cassini division are 20 K to 40 K cooler than the optically thin regions of the rings (Fig. 2). These results can be compared with earlier measurements, where, relative to their lit sides, the unlit A and B rings are ~25 K cooler in Voyager data (*19*, *20*) and ~30 K cooler in Pioneer 11 data (*21*). The difference is probably seasonal, attributable to the lower solar elevation angle during the Voyager and Pioneer flybys.

The anticorrelation between optical depth and temperature may not seem surprising at first, but it should be noted that this result applies even in the C ring, where the optical depth is small, typically 0.1 to 0.4. At such low optical depths, most of the particles are directly illuminated by the Sun, and most particles transit the ring plane twice each orbit, which physically carries their heat from the lit side to the unlit side. The small increase in mutual shadowing at optical depth 0.4 appears sufficient to present a colder ring to the CIRS instrument. In addition, more frequent collisions in the denser portions of the rings may hinder the warmer, sunlit particles from crossing the ring plane.

CIRS observed both the lit side (at a phase angle of ~130°) and the unlit side (at 59° phase) of the A ring. The derived temperatures are comparable, which appears to contradict our expectation that the lit side of the rings should be the warmer. This is also inconsistent with Voyager and Pioneer temperature measurements, where the A ring's unlit side was colder than the lit side (*20*, *21*). The difference in phase angle may provide the explanation; when we viewed the lit side of the rings, the phase angle was significantly higher, so we were viewing mainly the dark sides of the ring particles. If the particles have low thermal inertia (*22*) and rotate slowly, then viewing the lit rings at high phase may be similar to viewing the rings' unlit face; in both cases, we are primarily seeing the cold side of the ring particles. A superficially simpler interpretation is that the A ring's particles have nearly isothermal surfaces, so that the temperature will be the same under any viewing geometry; however, this is contradicted by the Voyager and Pioneer 11 results and appears to be inconsistent with the temperature variations elsewhere in the rings.

On 11 June 2004, CIRS obtained numerous observations of thermal emission from Phoebe's surface (Fig. 3). In addition to distant disk-integrated observations on approach and departure, CIRS obtained three global and six regional maps of Phoebe, plus valuable local high-resolution observations near closest approach obtained simultaneously with the visible and infrared mapping spectrometer (VIMS) and the Cassini Imag-

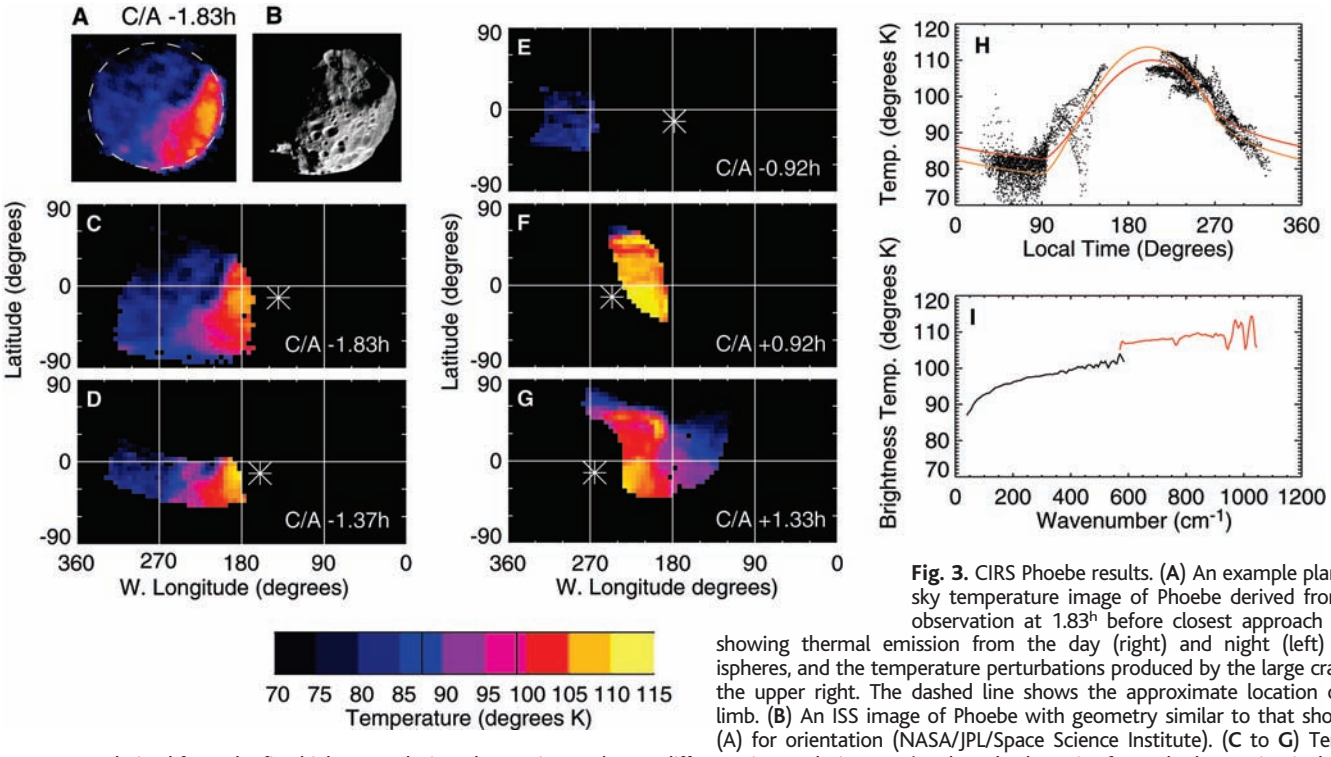


Fig. 3. CIRS Phoebe results. (**A**) An example plane-of-sky temperature image of Phoebe derived from the observation at 1.83^h before closest approach (C/A), showing thermal emission from the day (right) and night (left) hemispheres, and the temperature perturbations produced by the large crater in the upper right. The dashed line shows the approximate location of the limb. (**B**) An ISS image of Phoebe with geometry similar to that shown in (A) for orientation (NASA/JPL/Space Science Institute). (**C** to **G**) Temperature maps derived from the five highest resolution observations, taken at different times relative to C/A. The subsolar point for each observation is denoted with a star. (**H**) Low-latitude diurnal temperature variations derived from the midinfrared spectra, compared with one-dimensional homogeneous thermal models with thermal inertias of 2×10^4 (orange) and 3×10^4 (red) in units of erg cm^{-2} s$^{-1/2}$ K^{-1}. (**I**) A composite spectrum of low-latitude thermal emission at local times between 3 p.m. and 5 p.m., from the far-infrared channel (black) and midinfrared channel (red). Spatial coverage for the two detectors is slightly different, accounting for the slightly different temperatures. The discrete features in the spectrum are artifacts, but the trend of lower brightness temperatures at smaller wave numbers is real.

ing Science Subsystem (ISS) experiments (Fig. 3, A and B). Phoebe's relatively rapid rotation (its period is 9.1^h) allowed us to observe the same range of longitudes, including the prominent large crater seen in ISS images, over a wide range of local times during the flyby.

CIRS's midinfrared array pixels achieved a spatial resolution as small as 12 km for full-disk observations and 600 m near closest approach. Spectral resolution for most observations was 15.5 cm^{-1}. The spectra were indistinguishable from blackbodies at CIRS resolution and noise levels. For each map, we found the distribution of surface temperature (Fig. 3, C to G) as follows: We determined the thermal emission spectrum from each point on Phoebe's surface by averaging all overlapping spectra and fitting a blackbody curve to the mean 590 to 1100 cm^{-1} radiances. The pixel sensitivity allowed measurement of brightness temperatures as low as 75 K, and only the night-time polar regions of Phoebe were colder than this and thus not to be detected. There is strong topographic control of dayside temperature, particularly around the prominent large crater centered near 20°N, 220°W seen in Cassini visible-wavelength images (Fig. 3B), where temperatures on crater walls are up to 15 K warmer or colder than nearby horizontal surfaces, depending on illumination geometry. Low-latitude temperatures on Phoebe vary between 82 K before dawn to 112 K near the subsolar point. The diurnal variation can be matched with a bolometric albedo of 0.1, consistent with Voyager photometry (*23*), and a thermal inertia (defined as the square root of the product of thermal conductivity, density, and heat capacity) near 2.5×10^4 erg cm^{-2} s$^{-1/2}$ K^{-1} (Fig. 3H).

The CIRS far-infrared channel is not spatially coincident with the midinfrared array, and because of its relatively large aperture, its spatial resolution on Phoebe was never finer than 8 km. Expressed as brightness temperature, all far-infrared spectra show a roll-off toward low wave numbers that is characteristic of unresolved thermal inhomogeneities within the field of view (Fig. 3I). The principal sources of such inhomogeneities are probably variable illumination due to the strong topography mentioned above and a distribution of materials with different thermal inertias, such as blocks and fines. To investigate the diurnal behavior of Phoebe's surface temperature, we used only data taken within an hour of closest approach. We first estimated the total infrared flux by fitting the far-infrared spectrum with a combination of two blackbody spectra. We then used the single temperature that gave the same total flux as a representative surface temperature. At low latitudes, these temperatures varied between 72 K and 113 K throughout the day. From this diurnal behavior, we determined a mean thermal inertia near 2.0×10^4 erg cm^{-2} s$^{-1/2}$ K^{-1}, comparable to the value obtained from the midinfrared data.

Our thermal inertia values are about half of those derived from diurnal temperature variations on the Galilean satellites (*24*–*26*). This low thermal inertia, 1/40th that for solid water ice, implies a highly fragmented regolith within the top centimeter of Phoebe's surface. Phoebe's thermal inertia may be lower than that of the icy Galilean satellites, because its more rapid rotation results in a smaller thermal skin depth that samples the uppermost, most fragmented, portion of the regolith.

References and Notes

1. V. Kunde *et al.*, in *Cassini/Huygens: A Mission to the Saturnian Systems*, L. Horn, Ed., *SPIE Proc.* **2803,** 162 (1996).
2. F. M. Flasar *et al.*, *Space Sci. Rev.* **114**, 169 (2004).
3. A. P. Ingersoll, R. F. Beebe, B. J. Conrath, G. E. Hunt, in *Saturn*, T. Gehrels, M. S. Matthews, Eds. (Univ. of Arizona Press, Tucson, 1984), pp. 195–238.
4. C. D. Barnet, J. A. Westphal, R. F. Beebe, L. F. Huber, *Icarus* **100**, 499 (1992).
5. A. Sánchez-Lavega, S. Pérez-Hoyos, J. F. Rojas, R. Hues, R. G. French, *Nature* **423**, 623 (2003).
6. A. Sánchez-Lavega *et al.*, *Nature* **353**, 397 (1991).

7. R. F. Beebe, C. Barnet, P. V. Sada, A. S. Murrell, *Icarus* **95**, 163 (1992).
8. A. Sánchez-Lavega *et al.*, *Science* **271**, 631 (1996).
9. D. G. Andrews, J. R. Holton, C. B. Leovy, *Middle Atmosphere Dynamics* (Academic Press, New York, 1987), p.126.
10. B. J. Conrath, P. J. Gierasch, S. S. Leroy, *Icarus* **83**, 255 (1990).
11. B. Bézard, D. Gautier, *Icarus* **61**, 296 (1985).
12. B. J. Conrath, J. A. Pirraglia, *Icarus* **53**, 286 (1983).
13. B. Bézard, D. Gautier, B. Conrath, *Icarus* **60**, 274 (1984).
14. E. Anders, N. Grevesse, *Geochim. Cosmochim. Acta* **53**, 197 (1989).
15. M. H. Wong, P. R. Mahaffy, S. K. Atreya, H. B. Niemann, T. C. Owen, *Icarus* **171**, 153 (2004).
16. T. Owen, T. Encrenaz, *Space Sci. Rev.* **106**, 121 (2003).
17. H. Mizuno, *Prog. Theor. Phys.* **64**, 544 (1980).
18. J. B. Pollack *et al.*, *Icarus* **124**, 62 (1996).
19. R. Hanel *et al.*, *Science* **212**, 192 (1981).
20. R. Hanel *et al.*, *Science* **215**, 544 (1982).
21. L. Froidevaux, A. P. Ingersoll, *J. Geophys. Res.* **85**, 5929 (1980).
22. L. J. Spilker *et al.*, *Planet. Space Sci.* **51**, 929 (2003).
23. D. P. Simonelli *et al.*, *Icarus* **138**, 249 (1999).
24. J. R. Spencer, L. K. Tamppari, T. Z. Martin, L. D. Travis, *Science* **284**, 1514 (1999).
25. J. R. Spencer, thesis, University of Arizona, Tucson (1987).
26. J. A. Rathbun *et al.*, *Icarus* **169**, 127 (2004).
27. P. K. Seidelmann *et al.*, *Celest. Mech. Dynam. Astron.* **82**, 83 (2002).
28. A. Sánchez-Lavega, J. Rojas, P. Sada, *Icarus* **147**, 405 (2000).
29. L. W. Esposito, M. O'Callaghan, R. A. West, *Icarus* **56**, 439 (1983).
30. S. Albright, M. H. Elliott, and J. S. Tingley assisted with instrument commanding and data processing. D. Crick, M. de Cates, and S. Brooks assisted with observation designs. J. N. Cuzzi provided helpful comments on the manuscript. The authors acknowledge support from the NASA Cassini Project, the British Particle Physics and Astronomy Research Council, the Centre National d'Études Spatiales (CNES), and the Institut National des Sciences de l'Univers (CNRS/INSU).

Supporting Online Material
www.sciencemag.org/cgi/content/full/1105806/DC1
SOM Text
Figs. S1 and S2
Tables S1 and S2
References

29 September 2004; accepted 1 December 2004
Published online 23 December 2004;
10.1126/science.1105806
Include this information when citing this paper.

REPORT

Ultraviolet Imaging Spectroscopy Shows an Active Saturnian System

Larry W. Esposito,[1]* Joshua E. Colwell,[1] Kristopher Larsen,[1] William E. McClintock,[1] A. Ian F. Stewart,[1] Janet Tew Hallett,[2] Donald E. Shemansky,[2] Joseph M. Ajello,[3] Candice J. Hansen,[3] Amanda R. Hendrix,[3] Robert A. West,[3] H. Uwe Keller,[4] Axel Korth,[4] Wayne R. Pryor,[5] Ralf Reulke,[6] Yuk L. Yung[7]

Neutral oxygen in the saturnian system shows variability, and the total number of oxygen atoms peaks at 4×10^{34}. Saturn's aurora brightens in response to solar-wind forcing, and the auroral spectrum resembles Jupiter's. Phoebe's surface shows variable water-ice content, and the data indicate it originated in the outer solar system. Saturn's rings also show variable water abundance, with the purest ice in the outermost A ring. This radial variation is consistent with initially pure water ice bombarded by meteors, but smaller radial structures may indicate collisional transport and recent renewal events in the past 10^7 to 10^8 years.

The Cassini Ultraviolet Imaging Spectrograph (UVIS) (*1*) is part of the remote sensing payload of the NASA/European Space Agency (ESA) Cassini spacecraft. This spectrograph includes channels for extreme ultraviolet (EUV) and far ultraviolet (FUV) spectroscopic imaging, high-speed photometry of stellar occultations, solar EUV occultation, and a hydrogen/deuterium absorption cell (HDAC). UVIS science objectives are to study the composition, dynamics, and history of the saturnian system. UVIS extends the Voyager results of Broadfoot *et al*. in 1981 (*2*) and Sandel *et al*. in 1982 (*3*) through improved spectral and spatial resolution. We present results from the Saturn approach, systematic mosaics of the Saturn system, Phoebe, and Saturn Orbit Insertion (SOI). A highly structured and time-variable Saturn system is inferred from observations showing dynamic interactions between neutrals, ions, rings, moons, and meteoroids.

During the period preceding SOI, 25 December 2003 to 12 May 2004, spectral images of the saturnian system from the UVIS FUV and EUV spectrographs provided maps of the spatial distribution of emissions from neutral atomic hydrogen and oxygen. FUV mosaics of the Saturn system were obtained by stepping the slit position across the system in successive exposures of 200 s to 1000 s. Images in the atomic hydrogen Rydberg resonance transition at 121.57 nm (hydrogen Lyman-alpha) and the atomic oxygen resonance multiplet at 130.4 nm were obtained, both stimulated mainly in fluorescence of solar radiation. The observed oxygen has a full width at half maximum (FWHM) of ~4 Saturn radii (R_s) perpendicular to the orbital plane and ~16 R_S in the orbital plane (Fig. 1). Figure 1 is the second in the sequence of four images that follow the temporal variation of neutral oxygen. (The others are shown in figs. S1 to S4.) The figure shows a higher total abundance of oxygen that is more extensively distributed than in the previous and following images, which indicates that a transient event produced O in the Saturn system, followed by a rapid loss. The distribution of emission is asymmetric, with a peak near 3.7 R_S on the dark side of the planet. The total number of atomic oxygen atoms in this image is $> 4 \times 10^{34}$, or equivalent to 10^{12} g. Products of water dissociation, neutral oxygen, and OH [discovered in 1992 (*4*)] dominate the Saturn inner magnetosphere, in contrast to Jupiter, and H fills the entire magnetosphere, apparently extending through the magnetopause at far greater density than the ion population. The O and OH and a fraction of the H are products of water physical chemistry and derived ultimately from water ice.

The large abundance of neutrals limits the plasma abundance and electron temperature in the magnetosphere (*4*, *5*). The reactive system is self-limiting (*6*, *7*), because the ambient plasma ions are the source of neutral gas through reaction with E-ring grains. These, in turn, limit the plasma. With a water source, the system can achieve a quasi–steady state through the injection of energy supplied by hot electrons from the outer magnetospheric current system. One plausible source of water is collisions between larger, yet unseen parent bodies episodically resupplying water (*8*).

[1]University of Colorado, Laboratory for Atmospheric and Space Physics, 234 Innovation Drive, Boulder, CO 80303-7814, USA. [2]University of Southern California, 854 West 36th Place, Los Angeles, CA 90089–1191, USA. [3]Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, CA 91109–8099, USA. [4]Max-Planck-Institut für Sonnensystemforschung, Max-Planck-Strasse 2, D-37191 Katlenburg-Lindau, Germany. [5]Central Arizona College, 8470 North Overfield Road, Coolidge, AZ 85228–9778, USA. [6]Stuttgart University, Institute for Photogrammetry, Geschwister-Scholl-Strasse 24-D, 70174 Stuttgart, Germany. [7]California Institute of Technology, Division of Geological and Planetary Sciences, 1200 East California Boulevard, Pasadena, CA 91125, USA.

*To whom correspondence should be addressed. E-mail: larry.esposito@lasp.colorado.edu

The inferred change in total oxygen mass in the magnetosphere over about a 2-month period is $\sim 5 \times 10^{11}$ g, equivalent to the total estimated mass in the micrometer-sized particle component of the E-ring system (*9*). Over a period of 100 million years, the loss rate inferred from these observations would consume the entire E-ring mass, even including postulated parent bodies.

The observed oxygen variation could result from a single injection of neutral gas, which then dissipated within about 2 months. This rapid loss cannot be explained by satellite sweeping, but charge capture from the plasma ions mixed with the neutrals has a calculated time scale for removal of the neutral gas comparable to the observations. (*6*, *7*). In the charge-capture process, a neutral O atom loses an electron to a resident plasma ion trapped by the magnetic field of Saturn in the region of the magnetosphere between 3 and 5 R_S. The main reactions are

$$O + O^+ \rightarrow O^+ + O \quad \text{(Reaction 1)}$$

$$O + OH^+ \rightarrow O^+ + OH \quad \text{(Reaction 2)}$$

$$O + H^+ \rightarrow O^+ + H \quad \text{(Reaction 3)}$$

Half of these reactions produce a neutralized ion with escape velocity. Reactions 1 to 3 thus remove neutral gas from the system into the interplanetary medium while replacing old ions with new ions. The neutral gas lost must be replaced from icy material in the source region (*4*, *5*). The mass loss rates calculated from the UVIS observations (*6*, *7*) are now 10 times as large as those derived by Shemansky *et al*. (*4*). Jurac *et al*. (*8*) concluded that ion sputtering of the icy satellites could not provide the required mass input and that only the larger surface area provided by the micron-sized ring particles could replace water molecules rapidly enough. This leaves open how gas is extracted from the ring particles and the generation of new particles to replace those converted into the gaseous state. Shemansky *et al*. (*6*, *7*) propose the formation of water cluster ions through adsorption of ambient plasma ions on the grains, with subsequent dissociative recombination into the vacuum, thus supplying the gas to the system. An example of one reaction is given here.

$$O^+ + G \rightarrow O_2H_2^+ \cdot (H_2O)_n \cdot G \quad \text{(Reaction 4)}$$

$$O_2H_2^+ \cdot (H_2O)_n \cdot G + e \rightarrow O_2 + H_2 + Gx \quad \text{(Reaction 5)}$$

Reaction 4 represents the formation of a cluster ion through adsorption of an O^+ ion onto a grain surface (G) in which $O_2H_2^+$ is formed in a stabilizing bond with a cluster of n H_2O molecules in the solid. Reaction 5 is a recombination reaction of the cluster ion with an ambient plasma electron resulting in the release of O_2 and H_2 molecules into the vacuum. Reactions like 4 and 5 have been measured in the laboratory (*8*) in the gas phase. Reaction 5 is known to be among the most rapid transitions in ion physics and reaction 4 is considered to be the rate-limiting factor (*7*). In summary, reactions 1 to 3 remove mass from the Saturn system but leave the number of ions in the plasma unchanged, and reactions such as 4 and 5 extract neutral gas from the icy grain population, replacing the lost neutrals and removing plasma from the volume. Balance requires the generation of new ice grains and energetic plasma ions to produce water molecules. We propose that the local ionization process and energy required to maintain this system is produced primarily from the injection of energetic electrons from the outer magnetosphere (*7*).

The UVIS data show that the neutral gas in the magnetosphere is subject to transient phenomena that insert large amounts of icy ring material into the central plasma and deposit a large amount of energy into the plasma. Our results emphasize the completely different states of the Jupiter and Saturn magnetospheres: Saturn's neutral/ion mixing ratio is 30,000 times greater. The energy-deposition rate in the Saturn inner magnetosphere that is required to maintain the state of the system in the 3 to 6 R_S region (*5*) is estimated at 1.4×10^{10} W. This energy is mainly consumed in removing mass from the system. Energy deposition in the Jupiter magnetosphere, 3×10^{12} W, is primarily invested in radiative loss.

Previous UV images of Saturn's H_2 and H emissions from the Hubble Space Telescope

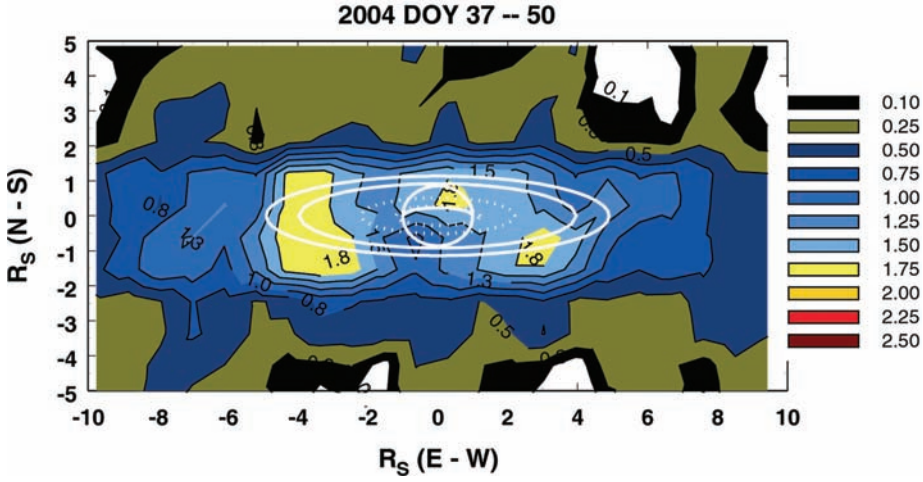


Fig. 1. The image of atomic oxygen emission from the Saturn system obtained in Cassini UVIS experiment observations from a range of 71 to 65 Mkm. The outline of the planet, orbits of Tethys, Enceladus (solid white lines), and outer A ring (dotted) are indicated. The subspacecraft point is at 13°4S latitude; solar radiation enters the system from the western side with a phase angle of 62° and a subsolar latitude of 23°S. The spatial resolution is 0.6 to 1.6 R_S. The image pixels are 0.9 × 0.9 R_S. North-south and east-west distances are indicated in R_S, with the origin at the center of the planet. Comparisons with UVIS observations made before and after this image show that it represents a much expanded emission distribution (figs. S1 to S4). Brightness values are in rayleighs. The expansion and subsequent decay of oxygen mass in the images show a characteristic time compatible with preliminary model calculations of charge-capture rates in the plasma.

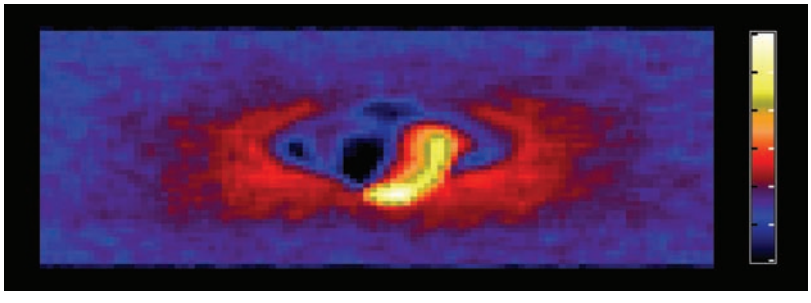


Fig. 2. Saturn's appearance in hydrogen Lyman-alpha. This image was acquired on 13 July 2004 from a range of 5 million km. Visible are the southern auroral zone, the sunlit crescent Saturn, the dark nightside, the dim rings, Saturn's shadow on the rings, and Saturn's extensive atomic hydrogen corona. The color bar is logarithmic, each tick representing a factor of 1.5 change in brightness.

(HST) show narrow auroral ovals in both the north and the south. These are probably due to primary electrons of 1 to 30 keV precipitating along the boundary of open and closed magnetic field lines (*10*). HST observed diurnal intensity variations of a factor of 10 and peak brightness near dawn, with evidence for auroral changes due to solar-wind variations (*11*–*13*). The UVIS image on 3 July 2004 (Fig. 2) shows strong UV polar aurora and dayglow emission at mid-latitudes. Magnetospheric neutral H also appears in an extended region near Saturn. The strongest emissions are from the south polar oval (~80° latitude). The north polar oval was tilted away from Cassini and appears only as polar limb emissions.

During July and August 2004, auroral emissions brightened episodically by up to a factor of 4. UVIS previously observed auroral brightenings at Jupiter due to solar-wind shock waves (*14*, *15*). A solar-wind shock wave on 25 July 2004 at 19:30 is followed by several days of enhanced solar-wind density in the Cassini Plasma Spectrometer (CAPS) data (*16*). This shock corresponds to the brightest UVIS-observed auroras so far. UVIS sees the two poles varying together. These Cassini UVIS and CAPS observations confirm that the solar wind perturbs Saturn's auroral emission.

Jupiter's spectrum from 2 January 2001 and Saturn's spectrum from 15 July 2004 are nearly identical spectrally (Fig. 3), showing H_2 band emissions from 115.0 to 170.0 nm (*17*). UVIS-derived hemispheric power input to Jupiter's auroral oval (10^{13} to 10^{14} W from electron precipitation) is about 200 times as large as that for Saturn's oval (10^{10} to 10^{11} W), in agreement with Voyager observations (*2*).

On 11 June 2004, as Cassini approached Saturn, the spacecraft flew by its outer moon, Phoebe. Distant observations of Phoebe were collected at a range of ≲380,000 km inbound and outbound, and mosaics were executed continuously through the closest approach at 2000 km. In addition, one full Phoebe rotation was monitored at 687,000 to 950,000 km. UVIS observed Phoebe in the EUV, FUV, and HDAC channels of the instrument.

The FUV spectrum (Fig. 4) was produced by averaging all pixels from the illuminated part of Phoebe during the mosaic taken near the closest approach (solar phase angle ~90°), divided by a solar spectrum obtained from the Solar Stellar Irradiation Comparison Experiment (SOLSTICE) (*18*) for the correct solar longitude. For comparison, the FUV reflectance spectrum of the Moon (similar solar phase angle) is also shown. The Phoebe spectrum displays a broad absorption feature centered near 160 nm, similar to that exhibited by H_2O frost and CO_2 frost (*19*, *20*). A model of Phoebe's reflectance spectrum combining H_2O frost, CO_2 frost, and a dark material to decrease the overall brightness is shown.

An FUV image of Phoebe obtained during the inbound portion of the flyby from a range of 31,300 km (Fig. 5) indicates that Phoebe is so dark that it blocks the Lyman-alpha emission of the Local Interstellar Medium (LISM), appearing silhouetted against the sky background. The heterogeneous surface reflectance shows brightness variations up to more than a factor of 2. These brightness variations are attributable to compositional variations (e.g., patchy ice distributions) across the surface, along with the rugged nature of Phoebe's surface, where steep slopes exposed to the sun appear brighter in this image.

As part of the investigation of Phoebe's origin, we searched for evidence of volatile emissions that would be indicative of comet-like activity such as that detected on Chiron. The set of long integrations acquired when Phoebe was ~1 pixel in size were summed to boost the potential signal level. The integrated spectrum showed no evidence of oxygen, carbon monoxide, or nitrogen emissions that would be anticipated if volatile activity were present. The 2σ upper limit for oxygen column density that UVIS could have detected given instrument sensitivity and the integration time is calculated to be $<2 \times 10^{13}$ cm^{-2} (*21*).

Phoebe has intrigued planetary scientists for decades because its great distance from Saturn (215 R_s) and Saturn's loose gravitational hold suggested that Phoebe was likely a captured object (*22*, *23*). Scientific curiosity has centered around the question of Phoebe's origin: Is it an errant asteroid or was it formed deep in the outer solar system? By detecting ice on Phoebe's surface, UVIS joins the other Cassini remote-sensing teams in concluding that Phoebe originated in the outer solar system.

During Cassini's orbit insertion on 30 June 2004, the spacecraft made its closest approach to Saturn's rings. During that period, the instrument slit was oriented roughly in the radial direction, and the detector was read out every 10 s, producing two approximately radial scans. During the scans, different parts of the detector sampled the same ring radius at different times as the spacecraft flew over the rings. We have summed all spectra from the same saturnocentric distance to provide a single spectral radial profile of the rings in the FUV. At any given radial position, about 60 spectra were recorded at sequential 10-s intervals, weighted by the amount of time spent and the fraction of the UVIS image pixel that is filled. Data were obtained in one scan for the outer C and inner B rings and in a

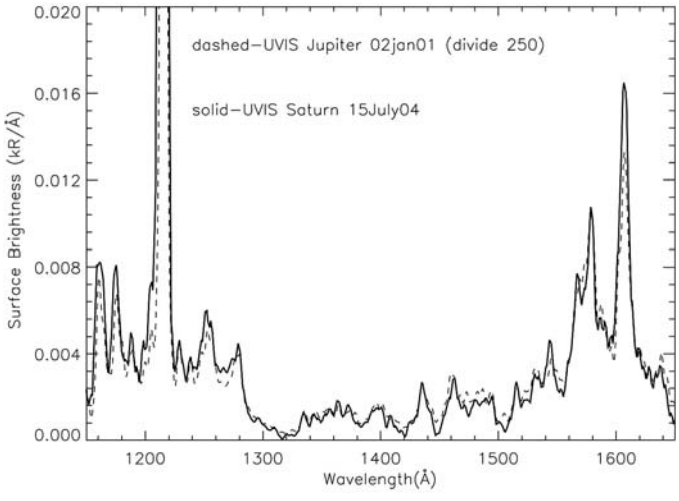


Fig. 3. Comparison of FUV spectra of Saturn (15 July 2004) and Jupiter (2 January 2001) southern auroral zones at ~0.5-nm resolution. Spectra are from a 28-min-duration system scan observation beginning at 03:38:22. Solar phase angle was near 90° for the dawn Saturn spectrum and 52° for the late afternoon Jupiter spectrum.

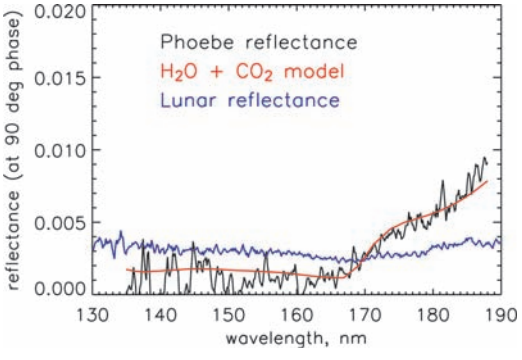


Fig. 4. Disk-averaged reflectance spectrum (obtained at 90° solar phase angle) of Phoebe compared with a model (red line), including laboratory spectra of H_2O frost and CO_2 frost (*20*). Also plotted is the UVIS-measured lunar FUV reflectance spectrum (blue line) for comparison. As a result of lower signal-to-noise ratios at wavelengths <150 nm, spectral features at the 3- to 6-nm scale may not be real.

second scan for the Cassini division and the A ring. Radial resolution is determined by the radial distance traversed by the footprint in the ring plane during each 10-s integration period, about 170 km in the first scan and 130 km in the second. To improve the photon-counting statistics, we adopted a resolution element of 150 km for both scans; this eliminates features that are below the noise level.

The ring brightness increases longward of 160 nm, consistent with absorption due to water ice at shorter wavelengths (*20*, *24*). At the shorter wavelengths, the ring signal is not distinguishable from the background except for the Lyman-alpha feature, which is a combination of scattered solar Lyman-alpha and Lyman-alpha from the local interstellar medium transmitted through the rings. Assuming that single scattering of sunlight dominates the observed brightness, which is a good approximation for this viewing geometry (*25*), we remove the geometric effects of the illumination and viewing angles and the optical depth to derive the product (AP) of the particle albedo, A, and the phase function, P (Fig. 6). In this approximation, fluctuations in AP indicate varying amounts of dark material mixed in with water ice. Without further observations, we cannot rule out more complex explanations of the fluctuations, for example, variations in vertical structure.

The A ring is the brightest ring region in the UV spectrum, the C ring is darkest, and the B ring is intermediate. The A-ring spectrum is similar to that of Phoebe (see Fig. 4), whereas the other rings are darker at long wavelengths, which indicates a lower fraction of water ice. In the A ring, the ice fraction, indicated in our data by UV reflectance longward of the water-ice absorption feature at 160 nm, increases outward to a maximum at the outer edge (Fig. 6). This large-scale variation is consistent with initially pure ice that has suffered meteoritic bombardment over the age of the solar system: The portions of Saturn's rings that have lower mass density become more polluted, and the subsequent transport of the darker material into the brighter rings over time explains the color gradients at the boundaries between rings (*26*).

We see variations over scales of 1000 to 3000 km in both ring scans that cannot be explained by this pollution mechanism operating over the age of the rings. These variations, which are not present in the Photopolarimeter Subsystem (PPS) optical depth profile, appear as small bumps and wiggles in the UVIS data in Fig. 6. Ballistic transport of spectrally neutral pollutants from meteoroids striking the rings produces broad gradients over this time period (*26*), too long to explain the finer structure in Fig. 6. We propose a class of smaller renewal events in which a small moon residing within the rings is shattered by an external impactor (*27*–*29*). We calculate the size of such a moon below. The interior of the moon has been shielded from external meteoritic bombardment and thus contains more pristine water ice. Because the amount of meteoroid pollution provides a rough clock to estimate the age of the rings (*26*), these random events reset that clock locally, making the material at that radial location younger and purer. As these purer ring particles collide with others, they interchange unconsolidated material on their surfaces, and the region of purer water-ice spectrum spreads radially, creating an ever-widening band of brighter material.

To estimate the time since such a reset event, we estimate the rate at which diffusion spreads the material. Using the kinematic viscosity $\nu \sim 280$ cm^2/s, the mass extinction coefficient $\kappa \sim 0.013$ cm^2/g (*30*, *31*), and the estimate of the fraction of regolith lost in a collision $f \sim 0.1$ (*32*), we calculate an effective diffusion coefficient for interchanging regolith material in the A ring

$$D_C \sim f * \nu \sim 30 \text{ cm}^2/\text{s}$$

Dimensional scaling gives an estimate of the time for such a renewal event to spread Δr = 1000 km as

$$T = \Delta r^2/D_C \sim 10^7 \text{year}$$

which is surely a lower estimate because f is uncertain. The mass needed to cover 10% of the surface area of average optical depth τ = 0.5, over an annulus Δr at saturnocentric distance r = 130,000 km is

$$M = 0.1 * 2\pi r\, \Delta r\, \tau/\kappa \sim 3 \times 10^{19} \text{g}$$

This mass is equivalent to a moon with radius $R \sim 20$ km.

Thus, the radial variations we interpret as due to differential pollution in our data set are consistent with the disruption of several small moons in the A ring in the past 10^7 to 10^8 years. This is one possible explanation of the variation in composition. Any process that releases pristine material, for example, the dredging of fresh material on ring particle surfaces by more energetic interparticle collisions at density wave locations, will create a brighter region in the rings. When the small moon Pan (*33*), $R \sim 10$ km, now residing in the nearby Encke gap, is eventually shattered by an external impact in the next 10 to 100 million years (*34*), our descendents will be treated to a spectacular sight: The gap will close up, and for some 10 to 100 million years thereafter a brighter radial swath of purer water ice at its former location will gradually spread and darken. Our interpretation of the spectral var-

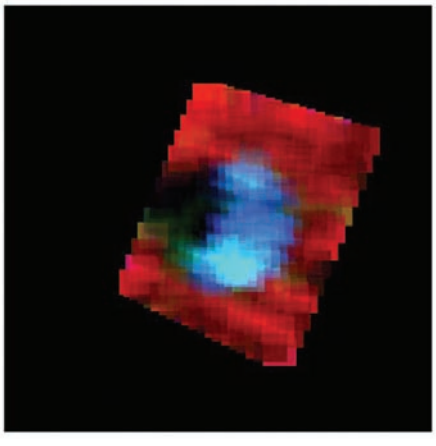


Fig. 5. FUV image of Phoebe, with solar illumination from the right. The red background corresponds to Lyman-alpha emission, blocked by Phoebe. Blue/green regions correspond to the longer FUV wavelengths (160 to 190 nm) and represent variability in brightness across the surface of Phoebe (up to a factor of more than 2 on the illuminated hemisphere).

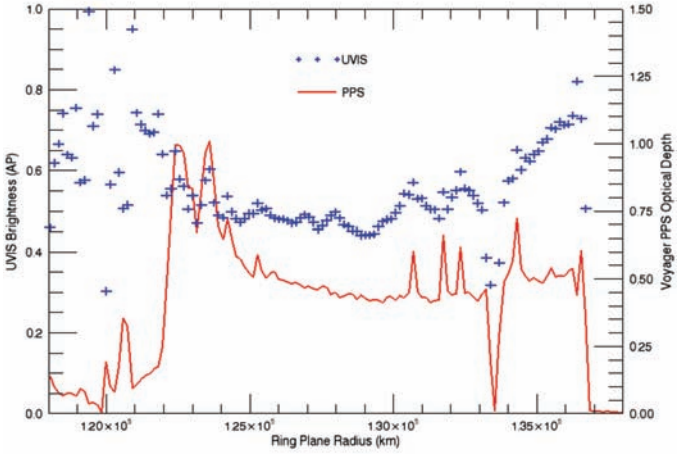


Fig. 6. The product of ring particle single-scattering albedo A, and particle phase function P, measured by UVIS, assuming that single scattering of sunlight dominates the measured brightness (blue). Absolute scale for AP is approximate. The Voyager PPS optical depth (red) shows the radial structure of the Cassini division (1.175 × 10^5 to 1.22 × 10^5 km) and the A ring. The UV brightness increases gradually toward the outer edge of the A ring with a sharp increase beyond the Encke gap at 1.335 × 10^5 km and a spike outside the Keeler gap at 1.365 × 10^5 km. Fluctuations at length scales of 1000 to 3000 km seen here and in the B ring indicate local variations in ring surface composition.

iations in Saturn's ring is thus consistent with a continual recycling of material between rings and moons, with the moons resupplying ring material when they are broken apart by collisions or meteorite impacts.

References and Notes

1. L. W. Esposito *et al.*, *Space Sci. Rev.*, in press.
2. A. L. Broadfoot *et al.*, *Science* **212**, 206 (1981).
3. B. Sandel *et al.*, *Science* **215**, 548 (1982).
4. D. E. Shemansky, P. Matheson, D. T. Hall, H.-Y. Hu, T. M. Tripp, *Nature* **363**, 329 (1993).
5. D. E. Shemansky, D. T. Hall, *J. Geophys. Res.* **97**, 4143 (1992).
6. D. E. Shemansky, Cassini UVIS team, paper presented at meeting of Committee on Space Research, Paris, France, 18 July 2004.
7. D. E. Shemansky *et al.*, in preparation.
8. S. R. Jurac, R. E. Johnson, J. D. Richardson, *Icarus* **149**, 384 (2001).
9. A. Juhasz, M. Horanyi, *J. Geophys. Res.* **107**, 1066 (2002).
10. J. T. Trauger *et al.*, *J. Geophys. Res.* **103**, 20237 (1998).
11. S. W. H. Cowley, E. J. Bunce, R. Prange, *Ann. Geophysicae* **22**, 1379 (2004).
12. J. C. Gerard *et al.*, *J. Geophys. Res.* **109**, A09207 (2004).
13. R. Prange *et al.*, *Nature*, in press.
14. D. A. Gurnett *et al.*, *Nature* **415**, 985 (2002).
15. W. R. Pryor *et al.*, in preparation.
16. D. T. Young *et al.*, *Science* **307**, 1262 (2005).
17. C. Jonin, X. Liu, J. M. Ajello, G. James, H. Abgrall, *Astrophys. J. Suppl.* **129**, 247 (2000).
18. W. E. McClintock, G. J. Rottman, T. N. Woods, *SPIE Proc.* **4135**, 225 (2000).
19. S. G. Warren, *Appl. Optics* **25**, 2650 (1986).
20. J. K. Wagner, B. W. Hapke, E. N. Wells, *Icarus* **69**, 14 (1987).
21. The gas detection threshold is based on a minimum 2σ detectable level of 10 counts above background, instrument sensitivity at 130.4 nm of 3.4 cts/kR-s, integration time of ~0 hours, and solar-wind values for the electron density and temperature at Saturn.
22. M. Cuk, J. A. Burns, *Icarus* **167**, 369 (2004).
23. J. B. Pollack, J. A. Burns, M. E. Tauber, *Icarus* **37**, 587 (1979).
24. R. Wagener, J. Caldwell, *ESA Special Publication SP-281* **1**, 85 (1988).
25. L. Dones, J. N. Cuzzi, M. R. Showalter, *Icarus* **105**, 184 (1993).
26. J. N. Cuzzi, P. R. Estrada, *Icarus* **132**, 1 (1998).
27. J. E. Colwell, L. W. Esposito, *J. Geophys. Res.* **98**, 7387 (1993).
28. J. M. Barbara, L. W. Esposito, *Icarus* **160**, 161 (2002).
29. L. W. Esposito, J. E. Colwell, paper presented at meeting of American Geophysical Union, San Francisco, CA, 8 December 2003.
30. L. W. Esposito, M. O. Callaghan, R. A. West, *Icarus* **56**, 439 (1983).
31. L. W. Esposito, *Icarus* **67**, 345 (1986).
32. R. M. Canup, L. W. Esposito, *Icarus* **119**, 427 (1996).
33. M. R. Showalter, *Nature* **351**, 709 (1991).
34. J. E. Colwell, L. W. Esposito, D. Bundy, *J. Geophys. Res.* **105**, 17589 (2000).

35. Thanks to H. Tollerud for assistance with data processing, L. Bloom for producing the manuscript, and J. Cuzzi and the anonymous reviewers for helpful comments. This work is one part of the Cassini UVIS investigation, supported by the NASA Jet Propulsion Laboratory Cassini mission.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1105606/DC1
Figs. S1 to S4

23 September 2004; accepted 2 December 2004
Published online 16 December 2004;
10.1126/science.1105606
Include this information when citing this paper.

REPORT

Radio and Plasma Wave Observations at Saturn from Cassini's Approach and First Orbit

D. A. Gurnett,[1]* W. S. Kurth,[1] G. B. Hospodarsky,[1] A. M. Persoon,[1] T. F. Averkamp,[1] B. Cecconi,[1] A. Lecacheux,[2] P. Zarka,[2] P. Canu,[3] N. Cornilleau-Wehrlin,[3] P. Galopeau,[3] A. Roux,[3] C. Harvey,[4] P. Louarn,[4] R. Bostrom,[5] G. Gustafsson,[5] J.-E. Wahlund,[5] M. D. Desch,[6] W. M. Farrell,[6] M. L. Kaiser,[6] K. Goetz,[7] P. J. Kellogg,[7] G. Fischer,[8] H.-P. Ladreiter,[8] H. Rucker,[8] H. Alleyne,[9] A. Pedersen[10]

We report data from the Cassini radio and plasma wave instrument during the approach and first orbit at Saturn. During the approach, radio emissions from Saturn showed that the radio rotation period is now 10 hours 45 minutes 45 ± 36 seconds, about 6 minutes longer than measured by Voyager in 1980 to 1981. In addition, many intense impulsive radio signals were detected from Saturn lightning during the approach and first orbit. Some of these have been linked to storm systems observed by the Cassini imaging instrument. Within the magnetosphere, whistler-mode auroral hiss emissions were observed near the rings, suggesting that a strong electrodynamic interaction is occurring in or near the rings.

Magnetized planets such as Saturn have many complicated radio and plasma wave phenomena. Here we present the first results from the Cassini Radio and Plasma Wave Science (RPWS) instrument (*1*) during the approach and first orbit around Saturn. The RPWS instrument is designed to measure the electric and magnetic fields of radio emissions and plasma waves across a broad range of frequencies, from 1 Hz to 16 MHz for electric fields and from 1 Hz to 12 kHz for magnetic fields. A Langmuir probe is also included to measure the density and temperature of the local plasma. Our observations are organized in the order in which the data were obtained, starting with radio emissions detected during the approach to Saturn, continuing through the region near the closest approach, and ending ~3 months after orbital insertion.

The Voyager 1 and 2 spacecraft first established that Saturn is an intense radio emitter. The primary component of this radiation is called Saturn kilometric radiation (SKR), because the peak intensities occur in the kilometer wavelength range (*2*), typically at frequencies from ~100 to 400 kHz. Just as with the other giant planets, the intensity of this radio emission is modulated by the rotation of the planet. During the Voyager flybys of Saturn in 1980 to 1981, the SKR modulation period was found to be 10 hours 39 min 24 ± 7 s (*3*). Because the charged particles responsible for the radio emission are controlled by the magnetic field, which is linked to the deep interior of the planet, and because the planet has no visible surface, this period has been widely adopted as the rotation period of Saturn (*4*). However, measurements by the Ulysses spacecraft (*5*) have shown that the radio rotation period is not constant.

This variability has now been confirmed by the Cassini observations. The RPWS first began to detect SKR in early April 2002, at a radial distance of ~2.5 astronomical units (AU) from Saturn. As the spacecraft approached Saturn, the signal strength gradually increased to the point that an accurate measurement of the SKR modulation period could be obtained. A normalized power spectrum of the fluctuations in the SKR intensity (Fig. 1) shows a sharp peak at a period of 10 hours 45 min 45 ± 36 s. Compared with the Voyager spectrum (Fig. 1), it

[1]Department of Physics and Astronomy, University of Iowa, Iowa City, IA 52242, USA. [2]Observatoire de Paris, 92195 Meudon, France. [3]Centre d'Etude des Environments Terrestre et Planétaires (CETP)–L'Institut Pierre-Simon La Place, 78140 Velizy, France. [4]Centre d'Etude Spatiale des Rayonnements–CNRS, 31028 Toulouse, France. [5]Swedish Institute of Space Physics, SE-751 21 Uppsala, Sweden. [6]NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA. [7]Department of Physics and Astronomy, University of Minnesota, Minneapolis, MN 55455, USA. [8]Austrian Academy of Sciences, Space Research Institute, A-8042 Graz, Austria. [9]Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 4DU, UK. [10]Department of Physics, University of Oslo, Blindern, N-0316 Oslo, Norway.

*To whom correspondence should be addressed. E-mail: donald-gurnett@uiowa.edu

is clear that a substantial shift has occurred in the radio rotation period over the 23 years since the Voyager measurements were obtained. The reasons for this shift are poorly understood. Because the magnetic moment of Saturn is aligned almost exactly along the rotational axis (*6*), one possibility is that rotational wobble induced by the tilt in the magnetic axis is simply not enough to control the rotational modulation. Other second-order rotational effects (*7*, *8*) can then become important, such as slippage of the magnetospheric plasma relative to the rotation of the upper atmosphere of Saturn, the period of which is known to vary with latitude and time (*9*). In sharp contrast, the radio rotation period of Jupiter, which has its magnetic axis tilted by 9.6° relative to its rotational axis, has been constant within a matter of seconds for more than 50 years. In addition to the radio rotation period, direction-finding measurement with the RPWS while viewing the southern polar regions confirmed that the SKR source is located over the southern auroral zone. This is consistent with constraints placed on the source location from the earlier Voyager measurements, which did not include a direction-finding capability.

During the Voyager 1 and 2 flybys of Saturn, numerous impulsive radio bursts, called Saturn electrostatic discharges (SEDs), were detected in the frequency range from 100 kHz to 40 MHz (*10*, *11*). Subsequently, it was shown that these signals are produced by lightning in the atmosphere of Saturn (*12*, *13*). Typical SED bursts had durations of ~40 ms and occurred in very regular episodes lasting several hours each, with a repetition period ranging from 10 hours 0 min to 10 hours 9 min (*14*). This period corresponded to the rotation period of clouds near the equator, which suggested that the lightning originated from a storm system that was being convected around the planet near the equator. A similar occurrence pattern was observed during both flybys, which were separated by ~9 months, suggesting the storm system was long-lived.

During the Cassini approach to Saturn, SED bursts were first detected on 22 July 2003 at a radial distance of 1.08 AU from Saturn. This SED episode lasted a few hours and was extremely intense, with peak intensities 10 to 30 times greater than those observed by Voyager 1, after correcting for radial distance. On the basis of observations of terrestrial lightning obtained during the Cassini flyby of Earth (*15*), the SEDs in this episode are estimated to be a factor of 10^6 more intense than comparable radio emissions from terrestrial lightning. Although a few weak SEDs were observed after this first episode, strong and easily identified SEDs were not observed again for nearly a year, until 13 July 2004, well after the 1 July 2004 orbit insertion. Since that time, strong SEDs have been observed on an irregular basis over a period of several months, with highly varying intensities and rates. However, none of these have intensities comparable to the very intense episode that occurred on 22 July 2003.

The number of SED bursts detected per Saturn rotation (Fig. 2) in the several-month period after orbital insertion varies considerably. Typical episodes were several hours long, with burst rates varying from a few per hour during periods of low activity to more than 100 per hour during very active periods. When active, the episodes tend to reappear every 10 to 11 hours as the planet repeatedly carries the storm system back around into view. Three different storm intervals have been identified, labeled A, B, and C (Fig. 2). Each is characterized by a distinctly different period that is thought to correspond to the atmospheric rotation period at the latitude where the storm system was located.

Previous evidence that SEDs are due to storm systems in Saturn's atmosphere has been indirect (*13*). However, comparisons of the SEDs detected by the RPWS and images of Saturn obtained by the Imaging Science Subsystem have resulted in visual evidence for a link between SEDs and storm systems in Saturn's atmosphere (*16*). In particular, the period and phase of the SEDs in intervals A and C match closely with distinctive cloud features seen in southern midlatitudes. The timing of the SED episodes always precedes the passage of the cloud feature across the central meridian with respect to Cassini. These observations suggest that the SED bursts are observed preferentially when the storm system is on the night side of Saturn, because the spacecraft was approximately over the dawn terminator during the system C observations. This relationship is expected, because radio signals from lightning can propagate through the ionosphere with less attenuation on the night side.

Previous studies of planetary magnetospheres, including Saturn (*17*, *18*), have shown that a variety of plasma waves are produced in the solar wind upstream of the planet. During the Cassini approach to Saturn, the first evidence of plasma waves associated with the planet occurred on 22 March 2004, when a burst of electron plasma oscillations was observed at a radial distance of 825 Saturn radii (1 R_S = 60,268 km). This large distance was possible because the dawn-side approach trajectory of Cassini provided a favorable geometry for electrons to stream along the solar wind magnetic field from the bow shock to the spacecraft. At Saturn's orbit, the solar wind magnetic field tends to lie in the ecliptic plane at an angle nearly perpendicular to the direction of the Sun.

Over the next few months, the electron plasma oscillations became more common and intense. Finally, at 0945 UT (Universal Time) on 28 June 2004, the spacecraft crossed the bow shock at a radial distance of 49.2 R_S. The bow shock crossing was identified in the RPWS data by an abrupt broadband burst of electrostatic noise. Subsequently, six more shock crossings were observed as the shock alternately moved back and forth over the spacecraft. These occurred at 1030, 1800, and 2000 UT on 28 June and on 0018, 0255, and 0535 UT on 29 June. The first crossing of the magnetopause

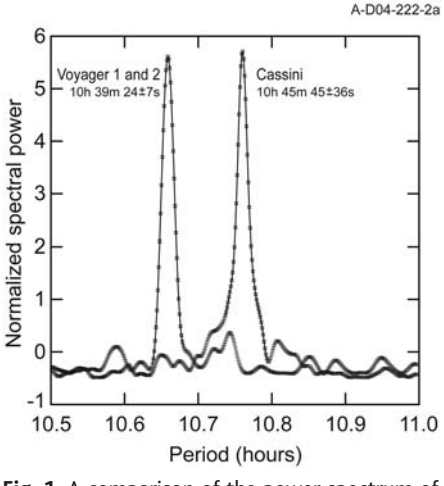


Fig. 1. A comparison of the power spectrum of variations in the SKR intensity as measured by Voyager and Cassini. The Cassini spectrum was accumulated over an interval of a little more than 1 year during the approach to Saturn, from 29 April 2003 to 10 June 2004. The radio modulation period measured by Cassini, 10 hours 45 min 45 ± 36 s, has shifted substantially from the Voyager era in 1980 to 1981.

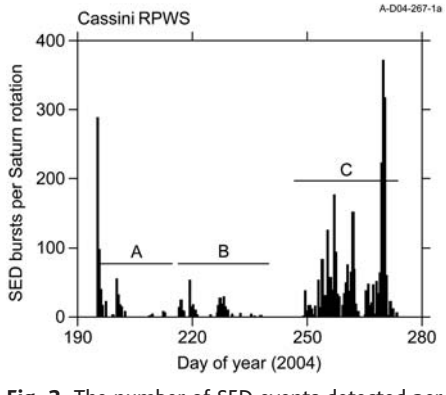


Fig. 2. The number of SED events detected per Saturn rotation from day 195 (13 July) to day 295 (1 October), 2004. Three distinct SED storm intervals can be identified, labeled A, B, and C, based on their periodicity. The periods of these storms were 10.72 hours, 10.26 hours, and 10.66 hours, respectively. Systems A and C have been associated with prominent cloud features in Saturn's atmosphere (*16*).

occurred at 1843 UT on 28 June, at a radial distance of 34.4 R_S. Numerous magnetopause crossings were then observed as the spacecraft approached Saturn. The last and final crossing into the magnetosphere was at 0244 UT on 29 June, at a radial distance of 30.6 R_S. The locations of these crossings are similar to those observed in 1979 during the outbound pass of the Pioneer 11 spacecraft (*19*), which occurred on the morning side of Saturn at a local time similar to the Cassini inbound trajectory.

Although several types of plasma waves were observed in the outer magnetosphere, as with the Voyager flybys (*17*, *18*), the strongest and most complex plasma wave emissions were in the inner region of the magnetosphere, inside of ~10 R_S. A frequency-time spectrogram of the electric field intensities detected by the RPWS in this region (Fig. 3) shows intense SKR from ~100 to 400 kHz. Narrowband emissions at slightly lower frequencies, such as the one starting at ~25 kHz at 2000 UT on 30 June, gradually drifting upward in frequency, and ending at ~100 kHz at 0100 UT on 1 July, are commonly observed in planetary magnetospheres (*20*) and are due to an electrostatic oscillation at the upper hybrid resonance (UHR) frequency. The UHR frequency is given by $f_{UHR} = \sqrt{f_c^2 + f_p^2}$, where f_c is the electron cyclotron frequency and f_p is the electron plasma frequency. The electron cyclotron frequency is given by $f_c = 28\ B$ Hz, where B is the magnetic field strength in nT: the electron plasma frequency is given by $f_p = 8980\sqrt{N_e}$ Hz, where N_e is the electron number density in cm^{-3}. The termination of the strong upper hybrid band at 0100 UT occurs near the inbound crossing of the outer edge of the A ring. After the closest approach to the planet, the upper hybrid band reappears at 0405 UT at a frequency of ~110 kHz, again near the outer edge of the A ring, and gradually drifts downward in frequency until it disappears at a frequency of ~30 kHz at 0900 UT. Because the UHR depends on f_c, it is useful to compare the frequency of the upper hybrid band with f_c. The f_c computed from the measured magnetic field strength (*21*) is well below the UHR frequency. This means that, except for a small correction, the upper hybrid band is essentially at the local f_p. A second, somewhat similar, narrowband emission can be seen at a frequency slightly above f_c on both the inbound pass, from ~0000 to 0100 UT, and on the outbound pass, from ~0405 to 0530 UT. This type of emission is also a common feature of planetary magnetospheres (*19*) and is an electrostatic electron cyclotron harmonic (ECH) wave. ECH waves occur near half-integral harmonics of f_c, i.e., near $(n + 1/2)f_c$, where n is an integer. The UHR and ECH waves are both driven by a loss-cone anisotropy in the trapped electron distribution and play an important role in the pitch-angle scattering and loss of energetic radiation-belt electrons. At even lower frequencies, after ~0400 UT on 1 July, several diffuse emissions were seen at frequencies below f_c. These are all whistler-mode emissions, because this is the only mode that can propagate at these frequencies. Whistler-mode emissions also play an important role in the pitch-angle scattering and loss of trapped radiation-belt electrons (*22*).

The electric field spectrum in the region where the spacecraft is passing over the rings, from ~0110 to 0409 UT, is complicated and not fully resolved (Fig. 3). The intense broadband noise, extending up to several tens of kHz from 0112 to 0248 UT, was produced by the firing of the rocket engine used to put the spacecraft into orbit around Saturn. A high-resolution frequency-time spectrogram (Fig. 4) of electric field waveforms obtained during the outbound pass over the rings from ~0309 to 0348 UT shows a broad V-shaped emission centered on ~0330 UT, with well-defined low- and high-frequency cutoffs that range from ~1 to 8 kHz. The V-shaped frequency-time shape of this emission is similar to a commonly occurring terrestrial whistler-mode emission called auroral hiss. At Earth, auroral hiss is known to be generated by magnetic field-aligned beams of low-energy (a few eV to several keV) electrons (*23*) associated with the current system responsible for the aurora. The V-shaped spectral feature arises from propagation along the whistler-mode resonance cone, which directs the radiation along a cone, the axis of which is aligned along the magnetic field. The cone angle increases with frequency, which accounts for the V-shaped frequency-time structure. If the electron plasma frequency is less than f_c, as is often the case in Earth's auroral regions and is the case here, then the upper cutoff frequency of auroral hiss is at the local f_p (*24*). Calculations using a ray path model show that the auroral hiss–like emissions detected by Cassini originate from a source close to the B ring, at a radial distance of ~1.7 ± 0.1 R_S. The existence of such auroral hiss–like emissions near Saturn's rings indicate that a low-energy beam of electrons is being accelerated outward away from the ring at a radial distance of

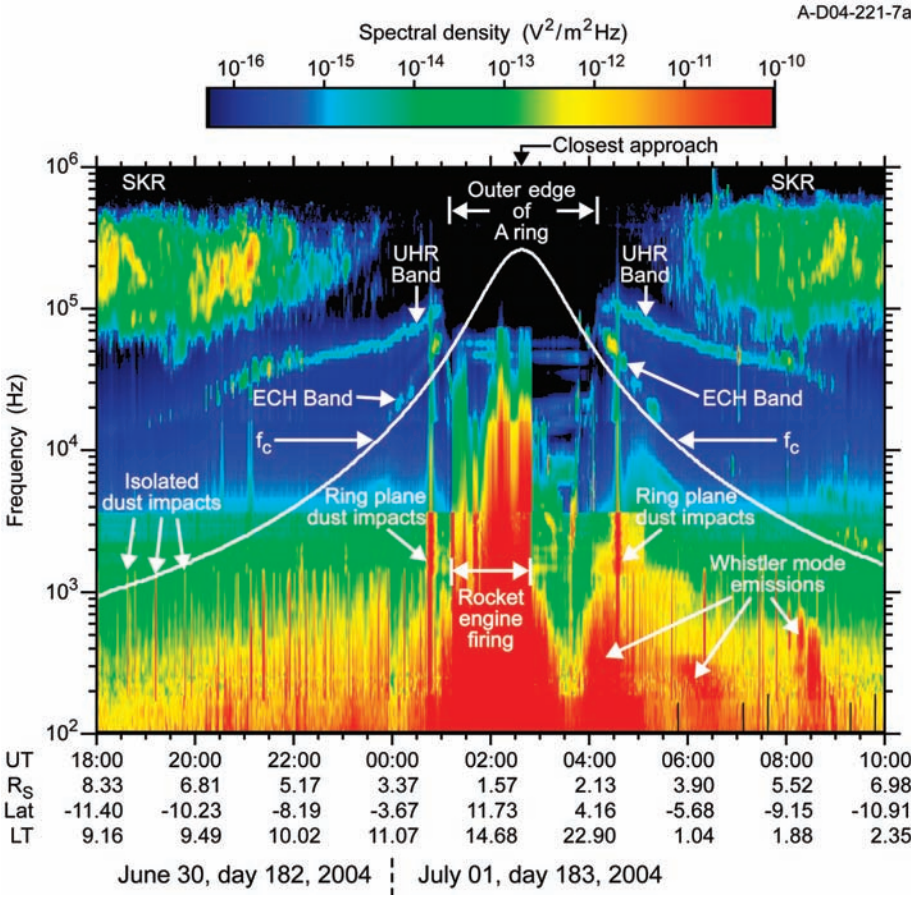


Fig. 3. A color frequency-time spectrogram showing the electric field intensities detected by the RPWS during the Cassini pass through the inner region of the saturnian magnetosphere. The intensities are coded according to the color scale at the top of the spectrogram. The white line is the f_c computed from Cassini magnetic field measurements (*21*). Lat., latitude; LT, local time.

~1.7 R_S, near the synchronous orbit point where the gravitational and centrifugal forces cancel. This electron beam is most likely associated with a current system induced by an electrodynamic interaction between the rings and the corotating magnetospheric plasma.

The spectrogram (Fig. 4) also shows a large number of narrowband emissions in the frequency range from ~6 to 11 kHz. Similar narrowband emissions were observed by Voyager 1 and 2 at Saturn (*17*, *18*) and are believed to be free-space (ordinary mode) electromagnetic waves generated by mode conversion from electron plasma oscillations, most likely in regions with steep density gradients. These narrowband emissions do not seem to penetrate into the region beyond the sharp upper cutoff of the auroral hiss–like emissions. Because free-space ordinary-mode electromagnetic waves cannot propagate at frequencies below f_p, this observation is consistent with the interpretation that the upper cutoff of the auroral hiss–like emission is at the local f_p.

Using the plasma wave resonances and cutoffs described above, we constructed an electron density profile through the inner region of Saturn's magnetosphere (Fig. 5). During the inbound pass, from ~2000 UT on 30 June to ~0100 UT on 1 July, and during the outbound pass from ~0405 to 0900 UT, the electron density was computed from the UHR bands (Fig. 3). The electron densities increase systematically with decreasing radial distance and reach a peak of a little more than 100 cm^{-3} near the outer edge of the A ring. Although not discernable in the spectrogram, weak UHR emissions allow the electron density to be extended into the region where the density decreases rapidly inside the outer edge of the A ring. Because of the noise from the rocket engine firing, no electron densities could be obtained from 0112 to 0147 UT. However, starting immediately after the termination of the rocket-engine firing, the electron density could be determined almost continuously from a combination of three measurements: the upper cutoff frequency of the auroral hiss–like emissions, the frequency of electron plasma oscillations, and the cutoff of the narrowband electromagnetic emissions, all of which are at f_p. The resulting profile shows that the electron density reaches a deep minimum of ~3×10^{-2} cm^{-3} at ~1.7 R_S, near the center of the V-shaped auroral hiss emission. Langmuir probe measurements, which give electron densities consistent with the plasma wave measurements, show that the electron temperature over the ring plane is low, only ~0.5 to 1.0 eV. In order to fit the ion current portion of the voltage sweep, ion masses in the range of 20 to 40 atomic mass units are required. The minimum in the electron density profile is close to the synchronous orbit point where the gravitational and centrifugal forces that act on the corotating magnetospheric plasma are in balance (*25*). Plasma generated inside this radius tends to be gravitationally attracted toward Saturn, whereas plasma generated outside this radius tends to be carried outward away from Saturn by interchange instabilities. Such dynamical processes may be able to account for the deep minimum in the plasma density at this point.

During the Voyager 2 ring plane crossing at Saturn, it was discovered that the plasma wave and radio astronomy instruments could detect the impact of small micrometer-sized dust particles (*18*, *26*). When a small particle strikes the spacecraft at a velocity exceeding a few kilometers per second, the particle is instantly vaporized and heated to a high temperature (10^4 to 10^5 K), thereby ionizing a substantial amount of the gas. As the resulting ionized gas cloud sweeps over the electric field antenna, some of the charge is collected by the antenna, thereby producing a voltage pulse into the receiver. Laboratory measurements show that the charge released, and therefore the amplitude of the voltage pulse, is proportional to the mass of the impacting particle.

During the pass through the outer regions of the saturnian system, the Cassini RPWS detected a small but nearly continuous level of dust impacts at a rate of a few impacts per minute. As the spacecraft approached the inbound ring plane crossing, the impact rate began to increase markedly ~2 min before the ring plane crossing, reached a peak at ~0046:33 UT, almost exactly at the ring plane crossing, and then decreased back to the pre-existing levels ~2 min later. Waveforms of the voltages on the *x*-axis electric dipole antenna (Fig. 6) near the time of peak impact rate show many very short pulses due to dust impacts. The peak impact rate at the ring plane crossing ranges from ~500 to 2000 impacts per second, depending on the counting threshold used. Similar dust impact waveforms were observed again during the outbound ring plane crossing at 0433:54 UT. The pulse amplitudes vary over a wide range (Fig. 6), from very small pulses that can just barely be detected to very large pulses that

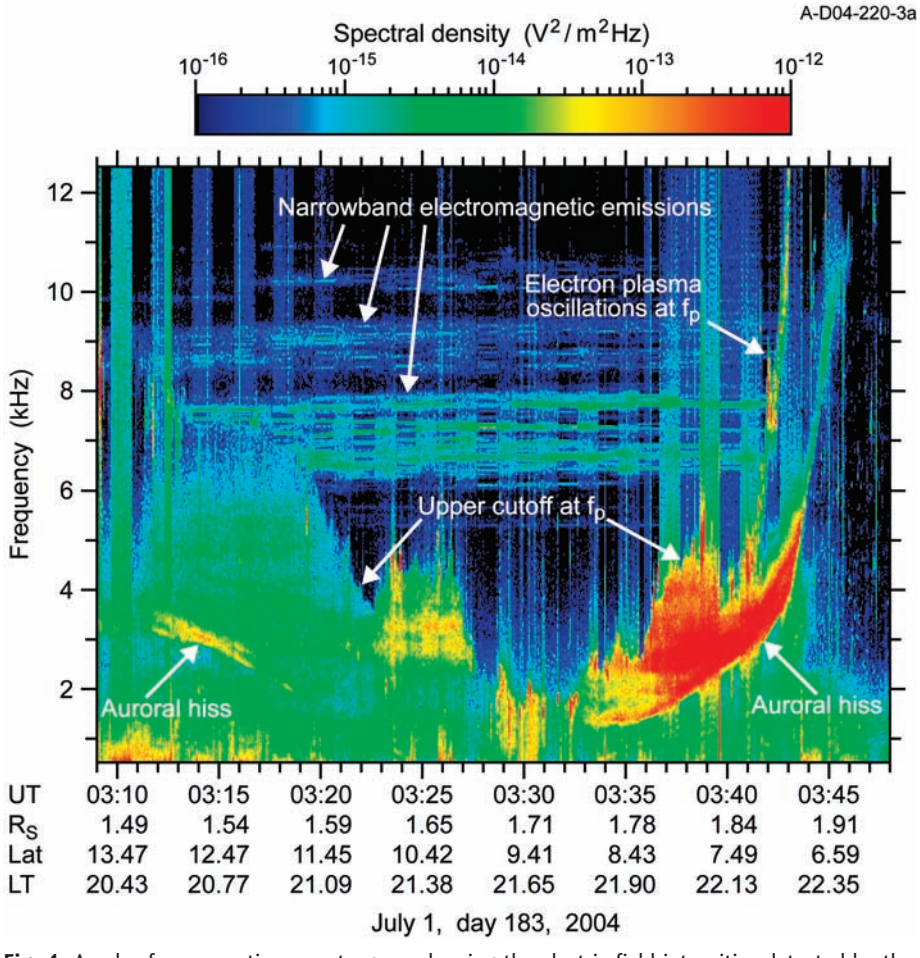


Fig. 4. A color frequency-time spectrogram showing the electric field intensities detected by the wideband waveform receiver over the rings during the outbound pass. The discontinuous changes in the background noise level every 30 minutes is due to cycling of the receiver bandwidth between 10 kHz and 80 kHz.

momentarily saturate the instrument. The amplitude of the pulses is controlled by two factors: the mass of the impacting particle and the proximity of the impact site to the electric antenna element. Both positive and negative pulses occur. The sign of the pulse is believed to be determined by which element of the dipole electric field antenna is nearest to the impact site. On the basis of previous studies of similar dust impacts (*27–31*), the masses of the particles are believed to be in the range from 10^{-12} to 10^{-9} g, corresponding to radii from ~0.5 to 5 μm, assuming a density of ~1 g cm^{-3}. Although the particle masses are difficult to estimate, the impact rate for any given pulse amplitude threshold can be very accurately determined. For both the inbound and outbound ring plane crossings, the north-south extent of the dust impact region was ~2000 km, with a north-south thickness at the half power-maximum-rate points of ~250 km, centered on the equator. From the pulse amplitude distribution, it appears that the largest particles, i.e., those with the largest pulse amplitudes, are more closely confined to the equatorial plane.

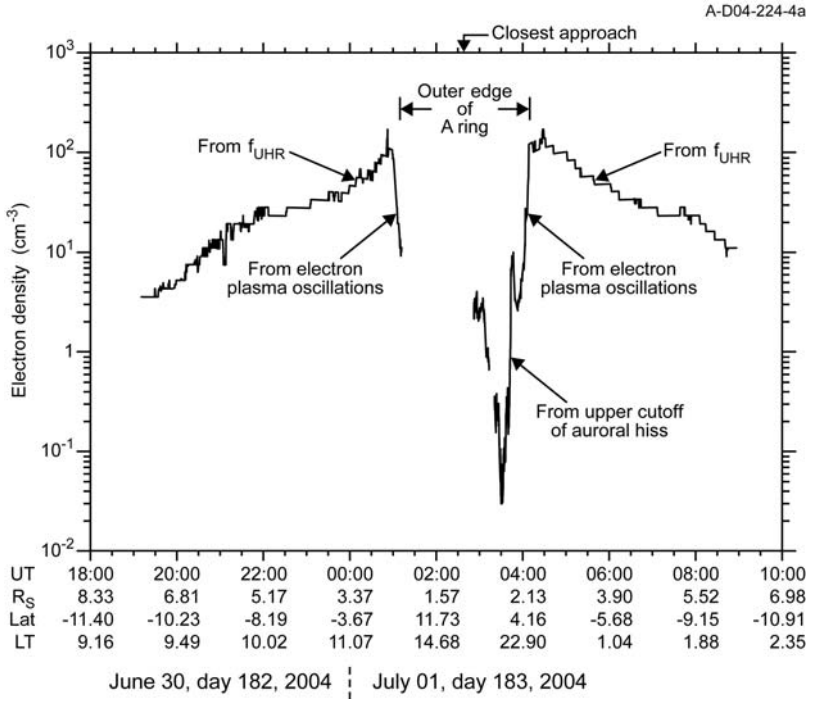


Fig. 5. The electron density profile through the inner region of the magnetosphere as determined from various plasma wave resonances and cutoffs. The sharp density depletion from 0100 and 0404 UT corresponds to the passage over the rings. f_{UHR} is the UHR frequency.

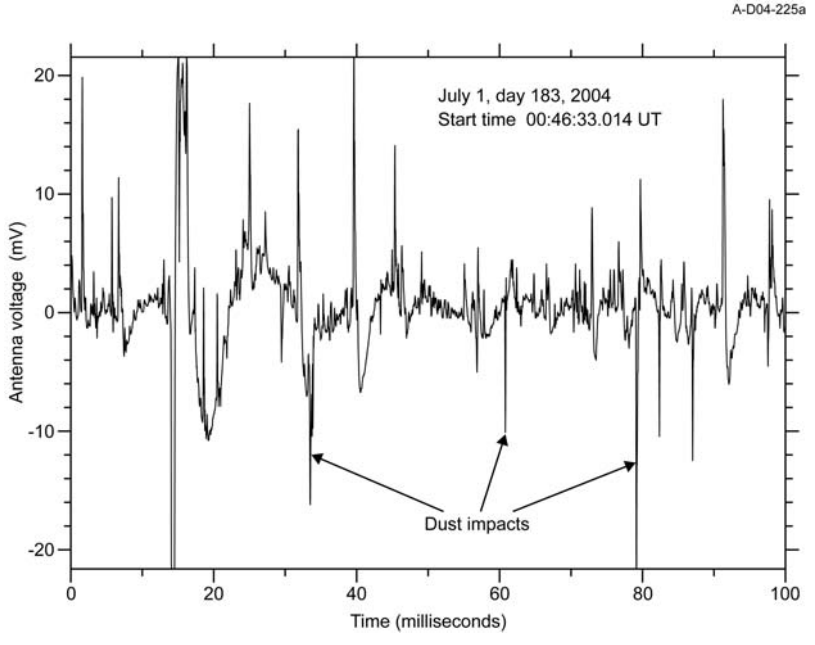


Fig. 6. The voltage waveform from the *x*-axis electric antenna during the inbound ring plane crossing. The short pulses with durations of a fraction of 1 ms are due to dust impacts on the spacecraft body.

References and Notes

1. D. A. Gurnett *et al.*, *Space Sci. Rev.* **114**, 395 (2004).
2. M. L. Kaiser, M. D. Desch, J. W. Warwick, J. B. Pearce, *Science* **209**, 1238 (1980).
3. M. L. Kaiser *et al.*, in *Saturn*, T. Gehrels, M. Shapley-Matthews, Eds. (Arizona Press, Tucson, 1984), pp. 378–415.
4. M. E. Davies *et al.*, *Celest. Mech. Dyn. Astron.* **63**, 127 (1996).
5. A. Lecacheux, P. H. M. Galopeau, M. Aubier, in *Planetary Radio Emissions IV*, H. Rucker, S. Bauer, A. Lecacheux, Eds. (Austrian Academy of Science, Vienna, 1997), pp. 313–325.
6. J. E. P. Connerney, L. Davis Jr., D. L. Chenette, in *Saturn*, T. Gehrels, M. Shapley-Matthews, Eds. (Arizona Press, Tucson, 1984), pp. 354–377.
7. P. H. M. Galopeau, P. Zarka, D. Le Queau, *J. Geophys. Res.* **100**, 26397 (1995).
8. P. H. M. Galopeau, A. Lecacheux, *J. Geophys. Res.* **105**, 13089 (2000).
9. A. Dessler, *Geophys. Res. Lett.* **12**, 299 (1985).
10. J. W. Warwick *et al.*, *Science* **212**, 239 (1981).
11. J. W. Warwick *et al.*, *Science* **215**, 582 (1982).
12. J. A. Burns, M. R. Showalter, J. N. Cuzzi, R. H. Durisen, *Icarus* **54**, 280 (1983).
13. M. L. Kaiser, J. E. P. Connerney, M. D. Desch, *Nature* **303**, 50 (1983).
14. P. Zarka, B. M. Pedersen, *J. Geophys. Res.* **88**, 9007 (1983).
15. D. A. Gurnett *et al.*, *Nature* **409**, 313 (2001).
16. C. C. Porco *et al.*, *Science* **307**, 1243 (2005).
17. D. A. Gurnett, W. S. Kurth, F. L. Scarf, *Science* **212**, 235 (1981).
18. F. L. Scarf, D. A. Gurnett, W. S. Kurth, R. L. Poynter, *Science* **215**, 587 (1982).
19. J. H. Wolfe *et al.*, *Science* **207**, 403 (1980).
20. W. S. Kurth, D. A. Gurnett, *J. Geophys. Res.* **96**, 18977 (1991).
21. M. Dougherty *et al.*, *Science* **307**, 1266 (2005).
22. C. F. Kennel, H. E. Petschek, *J. Geophys. Res.* **77**, 1 (1966).
23. D. A. Gurnett, in *High-Latitude Space Plasma Physics*, B. Hultquist, T. Hagfors, Eds. (Plenum, New York, 1983), pp. 355–375.
24. A. M. Persoon, D. A. Gurnett, S. D. Shawhan, *J. Geophys. Res.* **88**, 10123 (1983).
25. D. A. Mendis *et al.*, in *Saturn*, T. Gehrels, M. Shapley-Matthews, Eds. (Arizona Press, Tucson, 1984), pp. 546–589.
26. J. W. Warwick *et al.*, *Science* **215**, 582 (1982).
27. D. A. Gurnett *et al.*, *Icarus* **53**, 236 (1983).
28. M. G. Aubier, N. Meyer-Vernet, B. M. Pedersen, *Geophys. Res. Lett.* **10**, 5 (1983).
29. N. Meyer-Vernet *et al.*, *Icarus* **132**, 132 (1998).
30. P. Oberc, W. Parzydlo, O. L. Vaisberg, *Icarus* **86**, 314 (1990).
31. B. T. Tsurutani *et al.*, *Icarus* **167**, 89 (2004).

32. We thank the entire Cassini team at NASA Headquarters and the Jet Propulsion Laboratory (JPL) and the RPWS team at the University of Iowa for their outstanding support. The research at the University of Iowa was supported by NASA through contract no. 961152 with JPL, and the research at Observatoire de Paris and at CETP was supported by Centre National d'Etudes Spatiale.

17 September 2004; accepted 29 November 2004
Published online 16 December 2004;
10.1126/science.1105356
Include this information when citing this paper.

REPORT

Oxygen Ions Observed Near Saturn's A Ring

J. H. Waite Jr.,[1] T. E. Cravens,[2*] W.-H. Ip,[3] W. T. Kasprzak,[4] J. G. Luhmann,[5] R. L. McNutt,[6] H. B. Niemann,[4] R. V. Yelle,[7] I. Mueller-Wodarg,[8] S. A. Ledvina,[5] S. Scherer[1]

Ions were detected in the vicinity of Saturn's A ring by the Ion and Neutral Mass Spectrometer (INMS) instrument onboard the Cassini Orbiter during the spacecraft's passage over the rings. The INMS saw signatures of molecular and atomic oxygen ions and of protons, thus demonstrating the existence of an ionosphere associated with the A ring. A likely explanation for these ions is photoionization by solar ultraviolet radiation of neutral O_2 molecules associated with a tenuous ring atmosphere. INMS neutral measurements made during the ring encounter are dominated by a background signal.

A tenuous atmosphere in the vicinity of the saturnian rings was predicted to exist as a consequence of the sputtering of atoms and molecules from ring particles resulting from collisions with energetic ions, with micrometeoroids, or with solar photons (*1*, *2*), although the energetic ion fluxes observed over the rings were very small (*3*). The ring particles are thought to primarily consist of water ice (*4*), indicating that sputtered or photoproduced material should consist of water molecules or species formed from water molecules. In particular, it has been suggested that O_2 should be the main constituent of a ring-plane atmosphere, and O_2 is known to be a product of the radiation-induced decomposition of ice (*1*, *5*, *6*). A tenuous ring ionosphere was also predicted to be present at Saturn (*7*).

The INMS (*8*) has two different inlets: the open source, which allows ions or neutral species to enter the analyzer, and the closed source, in which neutral species enter an antechamber before being ionized. For open-source-ion (osi) mode, the incident ions are deflected by quadrupole switching lenses, set to appropriate voltages, and then guided to the radio-frequency quadrupole mass analyzer, which selects the mass-to-charge ratio. The ions are then detected with a secondary electron multiplier. For the closed-source-neutral (csn) or open-source-neutral (osn) measurements, electrons generated by heating of the W filaments are used to ionize some fraction of the neutral particles, and the resulting ions are then analyzed (*8*, *9*). The instrument is tuned to specific incident particle speeds, and this compensation speed was set to 15 km/s for the time period of 4:00:00 to 4:06:30 coordinated universal time (UTC) on 1 July 2004 (day 183 of the year). The compensation speed just before this time period was 20 km/s and just after this time period was 5 km/s. At 4:06:30 UTC, the instrument switched from osi to osn mode. The switching lens transmission is sensitive to speeds within ≈10% of this compensation speed (*8*, *9*) for both open-source modes. The open-source field of view (FOV) is close to 6° (full width). The instrument detects ions that are present in a small volume in velocity space centered at the compensation velocity (*10*).

A mass-time plot for the INMS open-source signal (Fig. 1) shows a high count rate for mass numbers 1, 16, and 32, which can be interpreted as a result of H^+, O^+, and O_2^+ ions, respectively. The count rate before 4:00 UTC was low. For O^+ and O_2^+ ions, the count rate was greatest between 4:00 and 4:01 UTC and between 4:02 and 4:06 UTC. At later times, the instrument switched to neutral mode. For H^+ ions, the count rate was highest between 4:02 and 4:06 UTC. During this time period, the spacecraft was about 7000 km above the outer part of the A ring at a radial distance of 2.2 Saturn radii (R_s). The peak O_2^+ count rate was ≈200 counts per second at 4:05 UTC. During the total of ≈2 s of the observing period from 4:00 to 4:06:30 UTC, devoted to measuring mass-32 ions, the total number of counts was about 150 (*11*).

The mass spectrum (Fig. 2) created from osi observations (Fig. 1) shows noticeable peaks at masses 1, 16, and 32, confirming the presence of H^+, O^+, and O_2^+, respectively. Other mass peaks with smaller signals relative to the statistical uncertainty also are evident (Fig. 2) at mass 17 (perhaps OH^+), mass 18 (perhaps H_2O^+), mass 23 (perhaps Na^+), and several mass numbers near 40 to 45 (perhaps CO_2^+ with poor velocity compensation).

The count-rate time variations (Fig. 1) could conceivably be due either to the spatial distribution of the ions (in this case, the scale size is on the order of 1000 km) or to the variations in INMS pointing, which resulted from changes in the spacecraft attitude (or a combination of these). If we assume the latter as a working hypothesis, then the INMS

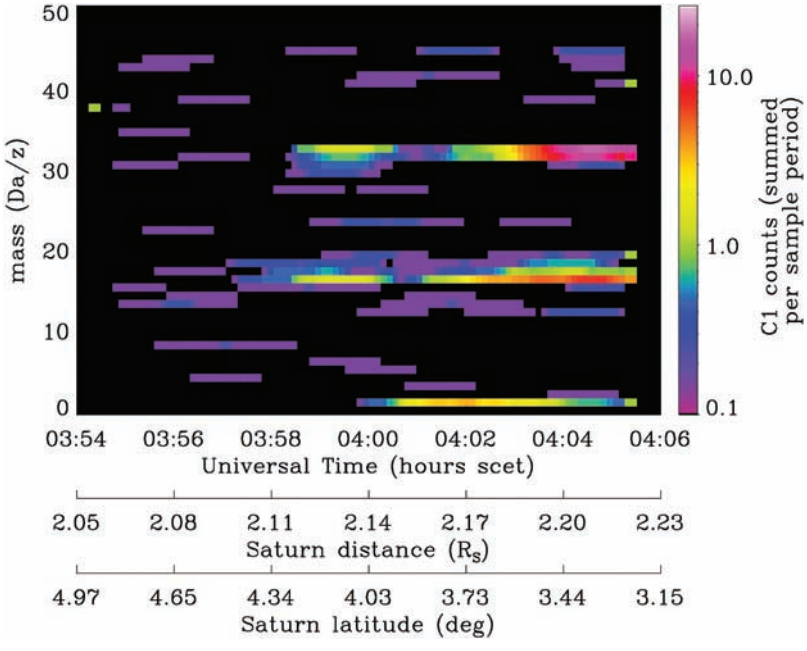


Fig. 1. Cassini INMS osi count rate (the color scale bar) versus mass number on the ordinate and Universal Time (scet) on the abscissa. Radial distance from the center of Saturn (in units of Saturn radii) is also shown. Da/Z, daltons per charge.

[1]Department of Atmospheric, Oceanic, and Space Physics, University of Michigan, Ann Arbor, MI 48109, USA. [2]Department of Physics and Astronomy, University of Kansas, Lawrence, KS 66045, USA. [3]Institute of Astronomy, National Central University, Chung-Li 32054, Taiwan. [4]NASA Goddard Spaceflight Center, Greenbelt, MD 20771, USA. [5]Space Science Laboratory, University of California, Berkeley, CA 94720, USA. [6]Applied Physics Laboratory, Johns Hopkins University, Laurel, MD 20723, USA. [7]Lunar and Planetary Laboratory, University of Arizona, Tucson, AZ 85721, USA. [8]Space and Atmospheric Physics Group, Imperial College, London SW7 2BW, UK.

*To whom correspondence should be addressed. E-mail: cravens@ku.edu

measurements can provide information on the ion distribution function over the A ring. The mass 16 and 32 data points used for Figs. 1 and 2 were averaged over 30-s intervals and placed in velocity space with the use of information on the INMS compensation velocity and pointing versus time. The INMS only observed a small region of velocity space (*10*), but within the observed region, the highest count rates were found for low vertical velocities (i.e., with respect to the ring plane) and for horizontal velocities (i.e., parallel to the ring plane) of about 5 km/s with respect to corotation. These measurements are consistent with either shell or ring distribution functions (*12*), but the existence of ions with smaller velocities with respect to corotation cannot be excluded on the basis of the INMS measurements. Indeed, measurements made by the Cassini Plasma Spectrometer (CAPS) instrument over the A and B rings (*13*) were consistent with co-rotating oxygen ions with a 0.75-eV temperature, suggesting the existence of ions with speeds closer to corotation than a few kilometers per second.

The count rate can be used to determine the ion flux into the INMS open-source inlet (*14*). We found that the average fluxes for masses 1, 16, and 32 (Fig. 2) are 3900 cm^{-2} s^{-1}, 900 cm^{-2} s^{-1}, and 2900 cm^{-2} s^{-1}, respectively. The peak count rates, and hence ion fluxes, for the time period near 4:04 UTC are about a factor of 2 higher than these average values (Fig. 1). The ion densities detected by the INMS near the compensation speed (*15*) can be found by dividing the fluxes by the compensation speed and are 0.0047 cm^{-3} (H^+), 0.0013 cm^{-3} (O^+), and 0.004 cm^{-3} (O_2^+). The peak densities are about twice these values and can also be used to determine peak ion distribution function values (*16*). The ratio of O^+ to O_2^+ is about 0.3, and the ratio of H^+ to O_2^+ is about 1.

Estimating absolute and total ion densities from the INMS data is difficult, because the INMS only sees a small fraction of the total ion velocity distribution. For a ring distribution function (*12*), the INMS would detect about 10% of the total distribution, whereas for a filled-in spherical distribution, INMS would detect about 1% of the total. Our estimated total ion number densities for these two assumptions are 0.1 and 1 cm^{-3}, respectively, for both H^+ and O_2^+ and 0.03 and 0.3 cm^{-3}, respectively, for O^+. The total electron density is then about $N_e \approx 0.2$ to 2 cm^{-3}. However, if a much colder component is also present, as suggested by CAPS (*13*), then the INMS might have observed an even smaller fraction of the total ion distribution than 1%, suggesting total densities greater than the above values. The electron density measured by the Cassini Radio and Plasma Wave Science (RPWS) experiment (*17*) increased from 6 to 25 cm^{-3} during the INMS observation period.

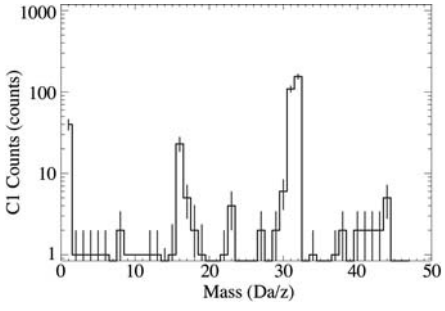


Fig. 2. INMS osi mode mass spectrum showing the number of counts versus mass number for times between 3:54:41 and 4:06:26 UTC. The 1σ statistical error is shown.

The INMS also operated in its csn mode during 3:54:41 and 4:06:26 UTC. The INMS is much less sensitive to neutral species because of the small fraction that are ionized in the instrument. A further complication was that the instrument cover had just been jettisoned about an hour earlier and some of the gas trapped in the instrument throughout the 7-year-long cruise phase of the mission was still present, resulting in a background count rate for all observable mass channels (*18*).

Photoionization of ring atmosphere neutrals by solar extreme ultraviolet or soft x-ray radiation seems the most likely explanation for the A-ring ions measured by the INMS, because the fluxes of energetic particles are negligible, probably because of absorption of the energetic plasma by the rings (*3*, *13*). The spacecraft was in Saturn's shadow at the time of the INMS observations; however, the outer A ring had been exposed to sunlight only an hour or so previously, which is considerably less than the ion lifetime (probably several hours, on the order of the ion bounce and mirror time). Cassini flew on the shady side of the rings; however, the optical depth of the A ring is ≈0.5 and considerable solar radiation can make it through the ring (*4*). Adopting a box model of the atmosphere-ionosphere environment over the ring in which ion production resulting from photoionization is balanced by transport loss of ions to the ring (after magnetic mirroring), we found the following approximate equation relating the ion density to the molecular oxygen density:

$$N(O_2^+)/N(O_2) \approx T_{bounce} I \approx 10^{-5} \quad (1)$$

where I is the ionization frequency for O_2 (*19*, *20*) and T_{bounce} is the ion lifetime (*21*). Similar relations for the H^+ and O^+ densities and the electron density can also be derived. The main source of O^+ ions is probably dissociative photoionization of O_2, although ionization of O could also make a contribution (*20*). Given our earlier estimate of the O_2^+ density $N(O_2^+) \approx 0.1$ to 1 cm^{-3}, Eq. 1 suggests that the neutral O_2 density is about 10^4 to 10^5 cm^{-3}, consistent with Ip's 3000 cm^{-3} O_2 density prediction (*1*). These densities are near the threshold of detection of the neutral mass spectrometer. The high values of the electron density measured by RPWS (*17*) imply either a neutral oxygen density substantially exceeding 10^5 cm^{-3} or an ion lifetime significantly greater than the bounce time.

The INMS, CAPS (*13*), and RPWS (*17*) instruments have revealed the existence of an atmosphere and an ionosphere/plasma in the vicinity of Saturn's A ring. The icy rings of Saturn generate a molecular oxygen-dominated atmosphere rather than one containing other water products (e.g., OH, OH^+, H_2O^+, and H_3O^+), which have shorter lifetimes as a result of sticking on the ring particle surfaces.

References and Notes

1. W.-H. Ip, *Icarus* **115**, 295 (1995).
2. G. R. Wilson, J. H. Waite Jr., *J. Geophys. Res.* **1989**, 17287 (1989).
3. S. M. Krimigis *et al.*, *Science* **307**, 1270 (2005).
4. L. W. Esposito *et al.*, in *Saturn*, T. Gehrels, M. S. Matthews, Eds. (Univ. of Arizona Press, Tucson, AZ, 1984), pp. 463–545.
5. R. E. Johnson, T. I. Quickenden, P. D. Cooper, A. J. McKinley, C. Freeman, *Astrobiology* **3**, 823 (2003).
6. R. E. Johnson, *Energetic Charged-Particle Interactions with Atmospheres and Surfaces* (Springer-Verlag, New York, 1990).
7. W.-H. Ip, *Nature* **302**, 599 (1983).
8. J. H. Waite *et al.*, *Space Sci. Rev.*, in press.
9. W. K. Kasprzak *et al.*, *Proc. SPIE* **2803**, 129 (1996).
10. The INMS in osi mode can measure ions in a velocity space volume approximately equal to the area of the FOV (at the compensation speed) multiplied by the length of compensation region (±1.5 km/s). The compensation speed was set to 15 km/s for 4:00 to 4:06:30 UTC and the volume was $\Delta V \approx 10^{16}$ $(cm/s)^3$.
11. The instrument electronics operation is such that for mass numbers greater than about 32 and for higher speeds, the velocity compensation cannot be perfectly achieved and there is a "spreading" of the signal to adjacent mass numbers. In particular, O_2^+ is seen at mass 31 as well as at 32. This is not a problem for lower mass numbers or for lower speeds, such as the 6-km/s ram speed (i.e., head-on gas speed) that the INMS will experience at Titan.
12. Ions created in the ring-plane environment from the ionization of neutral species moving at the local keplerian speed (≈15 km/s at the outer A ring) respond to the saturnian magnetic and corotation electric fields by both gyrating about the magnetic field and moving around the planet at the corotation speed (≈21 km/s at the outer A ring). In addition to this motion in the ring plane (which can be called horizontal), the ions should also have a small (≈1 km/s) velocity component along the magnetic field (approximately perpendicular to the ring plane, or vertical), which they inherit from the motion of the parent neutral atoms and molecules acquired during the surface ejection process and/or during gas interactions with the grains (*1*, *5*, *6*). The statistical distribution of these ions form a circle, or ring, in velocity space. This type of pick-up ion distribution is well known in other space environments (*22*, *23*). Near the outer A ring, the gyration speed of a newly formed ion in the co-rotating frame of reference is 5 to 6 km/s (i.e., the ring radius). The ring distribution can evolve into other distributions (such as a shell distribution) resulting from collisions or wave-particle interactions associated with magnetic fluctuations or waves (*22*).
13. D. T. Young *et al.*, *Science* **307**, 1262 (2005).
14. The INMS osi count rate, C', is given in terms of the flux of ions into the instrument and the incident ion energy by the relation: C' (counts per second) = $3.95 \times 10^{-3} + 5.35 \times 10^{-4}$ E (eV) (molecules cm^{-2} s^{-1}). The energy of an ion with a mass number of A (atomic mass units) is E (eV) = $0.0052Av^2$, where v is measured in kilometers per second. For A = 1, 16, and 32 and v = 15 km/s (the INMS compensation speed on this occasion), E = 1.2 eV, 19 eV, and 37 eV, respectively. We then found count rates per unit ion

flux for A = 1, 16, and 32 of 0.0046, 0.010, and 0.0240, respectively.
15. During its normal open-source (ions or neutrals) operation in the ionosphere and atmosphere of Titan, where the molecules are cold and the relative velocity is dominated by the spacecraft ram velocity, the entire distribution is within the instrument's compensation region volume (*10*), defined by the energy/velocity acceptance along the FOV and the velocity width perpendicular to this associated with the angular width of the FOV. However, during operations in the period after orbit insertion (over the A ring), only a small fraction of the ion distribution is captured in the compensation region.
16. The peak value of the distribution function, f, from INMS measurements for time periods when the compensation velocity is near the ring is the INMS density divided by the velocity space compensation volume (*10*). For H^+, O^+, and O_2^+, the peak f values are about 10^{-18} s^3/cm^6, 3×10^{-19} s^3/cm^6, and 10^{-18} s^3/cm^6, respectively.
17. D. A. Gurnett *et al.*, *Science* **307**, 1255 (2005).
18. The background rate of INMS in the csn mode has steadily diminished in the months after Saturn Orbit Insertion.
19. A. J. Kliore, D. P. Hinson, F. M. Flasar, A. F. Nagy, T. E. Cravens, *Science* **277**, 355 (1997).
20. W. F. Huebner, P. T. Giguere, *Astrophys. J.* **238**, 753 (1980).
21. Ions moving away from the ring plane along the magnetic field are reflected by the magnetic mirror formed by converging field lines and will return to the ring plane where they are removed from the plasma environment/ionosphere. This bounce time, T_{bounce}, is ≈3 hours and should also be the ion lifetime. However, because of a small northward offset of Saturn's magnetic dipole (*24*), some ions with very small vertical velocities might be trapped and have longer lifetimes.
22. A. J. Coates, *Geophys. Monogr. Am. Geophys. Union 61* (American Geophysical Union, Washington, DC, 1991), pp. 301–310.
23. J. G. Luhmann, in *Geophys. Monogr. Am. Geophys. Union 61* (American Geophysical Union, Washington, DC, 1991), pp. 5–16.
24. J. E. P. Connerney, L. Davis Jr., D. L. Chenette, in *Saturn*, T. Gehrels, M. S. Matthews, Eds. (Univ. of Arizona Press, Tucson, AZ, 1984), pp. 354–377.
25. We thank R. Johnson and F. Crary for useful and insightful discussions and comments. We are also very grateful for the hard work and dedication of our engineering, technical, and operational staff: G. Fletcher, E. Walter, R. Miller, J. Xu, J. Parajeko, D. Gell, and R. Thorpe. We thank D. Eddy, I. Robertson, and T. Hunt-Ward for technical help with the figures and text. Financial support from the NASA Cassini project is acknowledged.

28 September 2004; accepted 2 December 2004
10.1126/science.1105734

REPORT

Composition and Dynamics of Plasma in Saturn's Magnetosphere

D. T. Young,[1]* J.-J. Berthelier,[2] M. Blanc,[3] J. L. Burch,[1] S. Bolton,[4] A. J. Coates,[5] F. J. Crary,[1] R. Goldstein,[1] M. Grande,[6] T. W. Hill,[7] R. E. Johnson,[8] R. A. Baragiola,[8] V. Kelha,[9] D. J. McComas,[1] K. Mursula,[10] E. C. Sittler,[11] K. R. Svenes,[12] K. Szegö,[13] P. Tanskanen,[10] M. F. Thomsen,[14] S. Bakshi,[11] B. L. Barraclough,[14] Z. Bebesi,[13] D. Delapp,[14] M. W. Dunlop,[6] J. T. Gosling,[14] J. D. Furman,[1] L. K. Gilbert,[5] D. Glenn,[11] C. Holmlund,[9] J.-M. Illiano,[2] G. R. Lewis,[5] D. R. Linder,[5] S. Maurice,[3] H. J. McAndrews,[5] B. T. Narheim,[12] E. Pallier,[3] D. Reisenfeld,[14,15] A. M. Rymer,[5] H. T. Smith,[8] R. L. Tokar,[14] J. Vilppola,[10] C. Zinsmeyer[1]

During Cassini's initial orbit, we observed a dynamic magnetosphere composed primarily of a complex mixture of water-derived atomic and molecular ions. We have identified four distinct regions characterized by differences in both bulk plasma properties and ion composition. Protons are the dominant species outside about 9 R_S (where R_S is the radial distance from the center of Saturn), whereas inside, the plasma consists primarily of a corotating comet-like mix of water-derived ions with ~3% N^+. Over the A and B rings, we found an ionosphere in which O_2^+ and O^+ are dominant, which suggests the possible existence of a layer of O_2 gas similar to the atmospheres of Europa and Ganymede.

Most of what was known about Saturn's magnetosphere before Cassini's arrival was derived from the Pioneer 11 and Voyager 1 and 2 encounters from 1979 to 1981 (*1–5*) and from models based on that data (*6–8*). The measurements reported here were made with the Cassini plasma spectrometer (CAPS) (*9–11*) during the initial passage of the Cassini spacecraft through the near-equatorial regions of Saturn's magnetosphere. The CAPS instrument is made up of three plasma sensors.The first is the ion mass spectrometer (IMS), which measures ion energy per charge (E/Q) between 1 V and 50 kV with a resolution of $\Delta E/E = 0.17$. It simultaneously measures ion mass per charge (M/Q) from 1 to ~100 atomic mass units (amu) per charge, e, with a mass resolution $M/\Delta M \approx 60$. The second sensor is the electron spectrometer (ELS), which measures electron energy from 0.6 eV to 28 keV with $\Delta E/E$ = 0.17. The IMS and ELS are able to detect ion and electron densities as low as ~10^3 m^{-3}. The third sensor is the ion beam spectrometer (IBS), which measures ion E/Q with a very high resolution of $\Delta E/E = 0.017$, which is appropriate for narrowly beamed distributions.

The energy-time spectrogram (Fig. 1) and bulk plasma parameters (Fig. 2) give a broad overview of structures and events found 24 hours on either side of Cassini's closest approach to Saturn [02:39 universal time (UT) on 1 July 2004]. Within this time period on both inbound and outbound trajectories, we observed four regions with different physical and chemical characteristics: (i) The outer, or high-latitude, magnetosphere contains tenuous hot plasma dominated by H^+. (ii) A region we term the outer plasmasphere consists of highly variable, partially corotating plasma that contains a mixture of H^+, O^+, and water-group ions (denoted as W^+ and defined as a combination of OH^+, H_2O^+, and H_3O^+). (iii) The inner plasmasphere is less variable and closer to rigid corotation than the outer plasmasphere, and is made up primarily of O^+ and W^+. (iv) A layer of plasma consisting of O^+ and O_2^+ is located directly over the A and B rings. The boundaries separating these four regions are distinguished by changes not only in bulk plasma properties but also in chemical composition.

On the spacecraft's inbound trajectory, Saturn's magnetopause crossed over the spacecraft nine times between 34.6 and 30.6 R_S (1 R_S = 60,330 km). Just inside the magnetopause, plasma densities fell below the detection limits of the IMS. In particular, N^+, which might have been expected in the region of Titan's orbit (20.25 R_S), was not detected on either the inbound or outbound legs, possibly because the spacecraft was

[1]Southwest Research Institute, San Antonio, TX 78238, USA. [2]Centre d'Etude des Environnements Terrestre et Planétaires, Observatoire de St. Maur, 94107 St. Maur, France. [3]Observatoire Midi-Pyrenées, 31400 Toulouse, France. [4]Jet Propulsion Laboratory, Pasadena, CA 91109, USA. [5]University College London, Mullard Space Science Laboratory, Holmbury Saint Mary, Dorking, Surrey RH5 6NT, UK. [6]Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire OX11 0QX, UK. [7]Department of Physics and Astronomy, Rice University, Houston, TX 77251, USA. [8]Engineering Physics, University of Virginia, Charlottesville, VA 22904, USA. [9]VTT Industrial Systems, 02044, Finland. [10]Department of Physical Sciences, University of Oulu, 90014 Oulu, Finland. [11]Goddard Space Flight Center, Greenbelt, MD 20771, USA. [12]Division for Electronics, Norwegian Defense Research Establishment, N-2027 Kjeller, Norway. [13]KFKI Research Institute for Particle and Nuclear Physics, H-1525 Budapest, Hungary. [14]Space and Atmospheric Science Group, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [15]Department of Physics and Astronomy, University of Montana, Missoula, MT 59812, USA.

*To whom correspondence should be addressed. E-mail: dyoung@swri.edu

several ion-scale heights distant from Titan's orbital plane (*12*). At 14.4 R_S inbound and 13.6 R_S outbound (A and A′ in Figs. 1 and 2), Cassini crossed a ~0.2-R_S-thick boundary identifiable by an order-of-magnitude increase in density (from ~3 × 10^4 to ~3 × 10^5 m^{-3}). The density increase enabled us to establish that approximate plasma corotation occurs inside this boundary. Corotation is indicated by the broad peaks in count rates (Fig. 1) centered on the corotation energy (*13*) of protons and/or water-group ions. The A and A′ boundaries can be interpreted as a plasmapause, a structure that, at Earth, represents the last closed equipotential of plasma flow around the planet (*14*). At Saturn, however, magnetic flux tubes beyond the plasmapause are relatively empty because their **E** × **B** (where **E** and **B** are the planetary electric and magnetic fields, respectively) drift paths carry them through the magnetotail where centrifugal force can cause them to break open, spilling plasma down the tail (*15*).

Near ~9 R_S inbound and ~7.6 R_S outbound, Cassini crossed a second boundary (B and B′ in Figs. 1 and 2, B in Fig. 3) and moved into a region we identify as the inner plasmasphere. Here, ion motion is still organized by corotation. However, because the total ion energy is the sum of corotational and thermal energies, the bulk of the distribution appears above the theoretical corotation velocity curves in Figs. 1 and 2. A second characteristic of the inner plasmasphere is that the electron data are best fit by a bi-Maxwellian velocity distribution that yields separate densities and temperatures for two components. The colder component (~3 to 30 eV) increases in density and decreases in temperature with decreasing radial distance (Fig. 2). The hotter component (~100 to 1000 eV) has the opposite behavior; it decreases in density and increases in temperature with decreasing radial distance. Inside 6.3 R_S inbound (5.9 R_S outbound), the cold electron component is bounded by the proton corotation energy, a phenomenon that has yet to be explained (Fig. 1). Coincident with the change in bulk plasma characteristics in the inner plasmasphere is a change in ion composition. The abundance of O^+ and W^+ relative to protons increases sharply at the B boundary (Fig. 3) by more than an order of magnitude, reaching ratios greater than unity (because of pointing restrictions, there is no corresponding composition data outbound). Because the spacecraft is within ~1 ion-scale height of the equatorial plane, the change in composition must be caused by a radial gradient and not by scale-height effects. Neither the low relative abundance of O^+ and W^+ outside 9 R_S nor the steepness of the composition gradient agrees with Voyager observations (*6*, *16*) or with existing plasma generation models (*8*, *17*). These

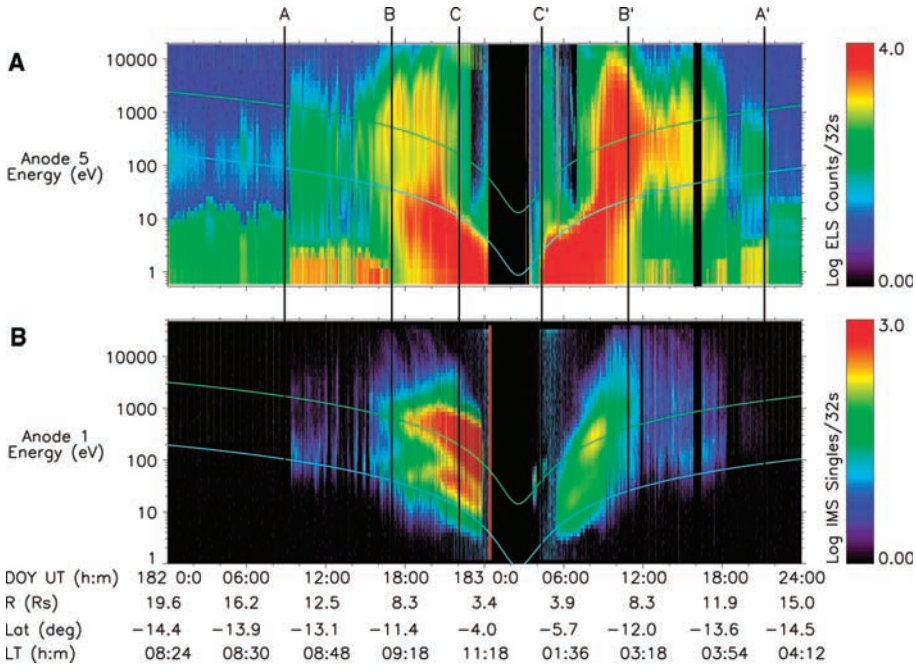


Fig. 1. An energy-time spectrogram showing particle counting rates (proportional to energy flux) in Saturn's magnetosphere for 30 June to 1 July 2004 (DOY 182 and 183). Data are averaged over 32 s. The color bars at right indicate rate intensity on a log scale. (**A**) Electron counting rates and (**B**) ion counting rates are shown. Energy increases along the vertical axes. Anodes 5 and 1 refer to the parts of the ELS and IMS detectors, respectively, that were most sensitive to electron and ion distributions over this period. The asymmetry in ion counting rates before and after closest approach (02:39 UT) is caused in part by a change in spacecraft pointing relative to the corotational flow direction. Universal time (UT), radial distance from the center of Saturn in R_S, latitude (Lat), and local time (LT) with respect to Saturn are shown on the horizontal axis. The curved lines superimposed on the plots give the energy corresponding to the theoretical corotation velocity (*13*) for O^+ (upper curve in both panels) and H^+ (lower curve). The data have been corrected for penetrating radiation background. Vertical black lines labeled A, B, and C indicate plasma boundaries on the inbound trajectory, whereas A′, B′, and C′ are the corresponding boundaries on the outbound leg. CAPS was turned off in the interval 00:00 to 03:25 UT on 1 July during the Cassini main engine burn.

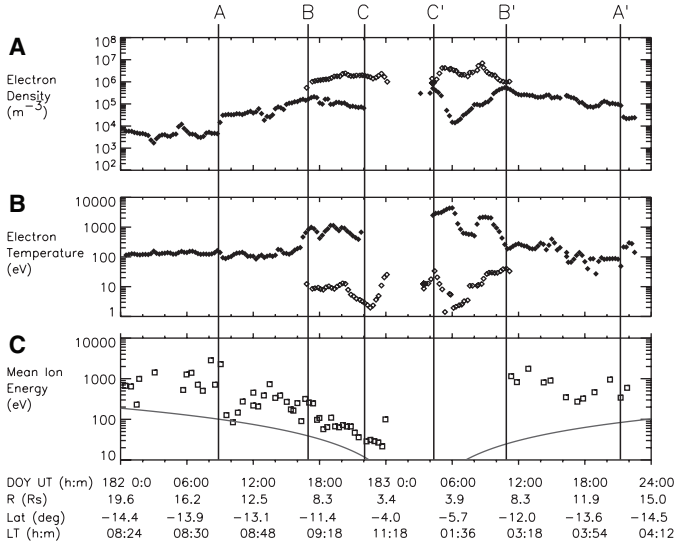


Fig. 2. Profiles of plasma parameters determined from moments of the electron and ion distributions. (**A**) Electron density. Solid symbols are the hot electron population. Open symbols are the cold electron population. The calculation of density is based on a fit to the electron distribution function assuming an isotropic bi-Maxwellian distribution of speeds in the spacecraft frame of reference. The two distributions become distinct and are well fit as a bi-Maxwellian only in the inner plasmasphere defined as the region between the B–C and B′–C′ boundaries. When the spacecraft is negatively charged, this density only includes the fraction of electrons with energies above the spacecraft potential. (**B**) Electron temperature derived in the same way as the densities shown in (A). (**C**) Proton mean kinetic energy is the statistical average energy of all observed protons and includes both the energy caused by the corotation of the plasma and thermal energy of the particles. On the outbound leg, poor spacecraft pointing relative to the corotational flow compromised the ion moment calculations. The curved line corresponds to the proton corotational energy.

results, although preliminary, could be attributable to changes in Saturn's magnetosphere after Voyager's observations or to differences in measurement methods.

Inside ~10 R_S, and as close as 4.4 R_S to Saturn, time-dispersed signatures of ions and electrons arriving at the spacecraft suggest that, at times, we are observing fresh plasma being injected deeper into the magnetosphere. Many of these features, which are too short to be visible at the scale of Fig. 1, are characterized by the initial arrival of high-energy ions (~10 keV) followed by the arrival of progressively lower energies, and then the appearance of low-energy electrons followed by progressively higher energies. Similar structures occur quite commonly at Earth (*18*) and have been observed at Jupiter (*19*), but have not previously been reported at Saturn. At Earth, the injections are caused by sudden increases in the global convection electric field. For rapidly rotating planets like Jupiter and Saturn, injections can be explained as centrifugally driven interchange motions in which cold dense plasma moves outward and is replaced by hotter, more tenuous plasma moving inward (*20*). In the corotating frame, gradient and curvature drifts caused by the magnetic field disperse ions and electrons eastward and westward, respectively, at a rate proportional to their thermal energies. The slope of energy versus time-of-arrival can be used to infer the age of the particles after injection. With this method, we estimated ages in the range 1.5 to tens of hours, indicating that the oldest injection signatures can persist for the order of one planetary rotation, which is comparable to the ages of similar events reported at Jupiter (*19*). The inner boundary of electron-injection signatures (which are the most prominent) is ~4.4 R_S. Consistent with interchange motions of magnetic flux tubes, a few instances of intense plasma acceleration were observed, which may have been caused by low-frequency magnetohydrodynamic (MHD) waves or strong Alfvenic perturbations of the magnetic field ($\Delta B/B > 0.1$) seen at the same time (*21–24*).

As Cassini proceeded deeper into the magnetosphere, the concentrations of O^+ and W^+ relative to H^+ remained near unity or above from ~8 R_S to ~5 R_S, then increased again at C and C′ (Figs. 1 to 3). We estimate that both N^+ and H_3O^+ are present at abundances of ~3% inward as far as 3.4 R_S (at that point the CAPS sensors were turned off as a precaution against degradation during the firing of the Cassini main engine). The N^+ ions, which are low energy (20 to 100 eV) (Fig. 2), cannot be ionized near Titan and transported inward because they would be heated to many keV in the process. We conclude that the N^+ observed inside 8 R_S, although its ultimate source may have been Titan's atmosphere (*25*), must be ionized locally. This could result from neutral nitrogen ejected from Titan that reaches the inner magnetosphere before being ionized (*12*), or ions from Titan that are implanted or adhere to the surfaces of the inner icy satellites or E-ring particles and subsequently are sputtered off (*26*). A possibility is that the nitrogen ions may be indicative of a small abundance of ammonia in these surfaces, an idea that has been proposed to explain morphologic features on Enceladus (*27*).

The inner plasmasphere can be identified with the E ring (~3 to 8 R_S), which is a rich source of water vapor (*28*). Because the ion population there has a low mean energy (Fig. 2C), it is also a region where the relative speeds between neutral gas moving at Keplerian velocities and ion species moving at corotation velocities are low, so that ion-molecule reactions become important (*29*). Our observation of H_3O^+ is interesting as an example of these reactions because it shows that hydronium ions are created in ion-molecule reactions that take place in a water-dominated atmosphere. One source is the reaction $H_2O^+ + H_2O \rightarrow H_3O^+ +$ OH, which, in addition to producing the ion, would also contribute to the observed OH torus (*30*). The production of H_3O^+ is similar to processes occurring in the water-rich environment of cometary comas as observed, for example, at comet Halley (*31*). The C and C′ boundaries also mark the point at which the suprathermal electron fluxes drop out (Fig. 2), possibly because of ionizing collisions with the neutral gas cloud associated with the icy satellites and E ring, which is densest at ~5 R_S (*32*).

When CAPS was turned back on at 03:25 UT after the main engine burn, the spacecraft was outbound at 1.65 R_S from the center of Saturn and 18,000 km above the inner part of the B ring. Between 03:37 and 04:09 UT, Cassini was in eclipse behind Saturn; however, the rings were still illuminated from the opposite side of the ring plane on magnetic field lines conjugate to the spacecraft. During this time, we observed a layer of ions (effectively an ionosphere) over the rings (Fig. 4). Based on Langmuir probe measurements, electron densities were $<3 \times 10^6$ m^{-3} (*33*), from which we conclude that the spacecraft very likely was negatively charged by a few volts, which prevented the ELS from entirely measuring the electron velocity distribution. A negative spacecraft potential would also attract positive ions and accelerate them into the IMS. This would improve the chances of measuring atomic or molecular hydrogen ions (corotational energy ~2 eV), which, nonetheless, are not evident (Fig. 4).

Electron count rates are lowest above the B ring, highest over the Cassini division, and intermediate above the A ring; in other words, they vary inversely with ring optical depth (*34*). This inverse correlation suggests that the electrons may originate from photoemission of magnetically conjugate particles in sunlight on the far side of the rings. The peak ion rates began ~5 minutes before the spacecraft reached the Cassini division, whereas the peak electron rates were highest directly over it (Fig. 4, A and B), possibly reflecting the effect of a negative spacecraft potential on electrons. The bulk of the ring ionosphere is made up of $M/Q = 16$ and 32 amu/e ions (Fig. 4C), which we identify as O^+ and O_2^+, respectively (*35*). This finding is consistent with corotation and a temperature of 0.75 eV for both species. Any hydrogen ions present (atomic or molecular) had relative concentrations <3%. After 03:57 UT, the spacecraft turned so that the IMS was no longer pointed in the corotation ram direction, causing ion count rates to drop (Fig. 4B). However, the electron distribution was closer to isotropy in the field-of-view of the ELS, permitting measurements to continue.

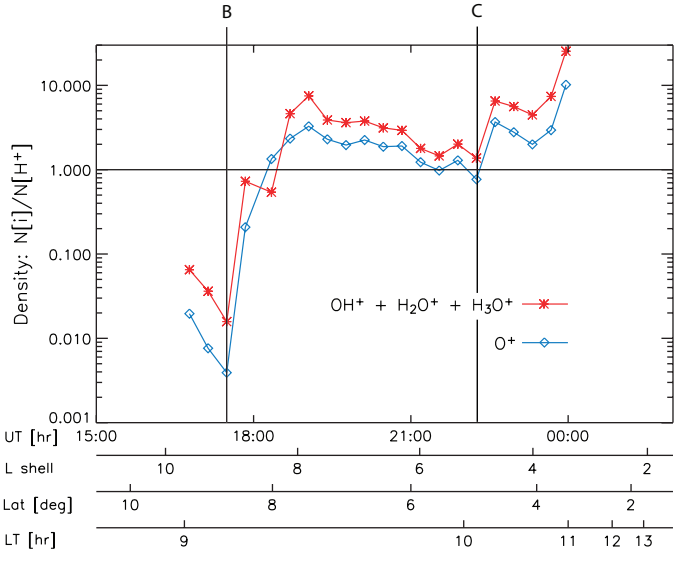


Fig. 3. The concentrations of O^+ and water-group ions are shown relative to H^+ on the inbound leg. Data are 20-min averages calculated by assuming all ions have an isotropic distribution in the spacecraft frame of reference. This assumption is not needed for Fig. 2C. Errors in relative density are estimated to be ±30% for O^+ and OH^+ and ±60% for H_2O^+ and H_3O^+. The L shell is equivalent to the radial distance R_S.

The detection of O_2^+ as a principal ion points to the existence of a layer of molecular oxygen gas over the rings, which we refer to as an atmosphere. Although O_2^+ can originate through photolytic processes in water vapor (*36*), the O_2 atmosphere over the rings, like the atmospheres of Europa and Ganymede, is more likely the result of radiation-induced decomposition of ice (*37*, *38*). In the inner plasmasphere, O_2 can be produced from ice by plasma with energy above a few keV. However, over the rings, energetic plasma fluxes are negligible, leaving production from ice by ultraviolet photons as the primary O_2 source (*39*). Although this is an inefficient process, the O_2 does not stick to the ring particles, whereas O and H_2O and its dissociation fragments stick or react at these temperatures. Therefore, O_2 is the dominant gas-phase species leading to the formation of O_2^+ and O^+ by photoionization in a ratio of ~4:1, which is close to our measurement (Fig. 4C) except for the period 03:39 to 03:43 UT. By using the photoproduction rates of O_2 from ice (*40*) as a lower limit to the source rate, together with photoionization and photodissociation loss rates, we estimate that the O_2 column above the rings is $>2 \times 10^{17}$ m^{-2}. O_2^+ ions that are produced north of the ring plane will oscillate about the magnetic equator, which lies ~2000 km north of the ring plane. These ions are eventually lost to the rings or to molecular collisions in one bounce period, giving an average O_2^+ density of $\sim 3 \times 10^6$ m^{-3}, which is consistent with our measurements. In isolation, photolysis of ice would also produce an H_2 atmosphere over the rings. However, the low concentration of hydrogen ions indicates that the surfaces of the ring particles, together with the ring atmosphere, must be treated as a complete system. For example, processes such as charge exchange and photodissociation of water or H_2 molecules can liberate hydrogen with sufficient energy to remove it from the rings, leaving behind an oxygen-rich atmosphere.

We have observed important features in the plasma distributions, dynamics, and composition of Saturn's magnetosphere that challenge current models. The data taken during Cassini's passage through the inner plasmasphere and its only passage through the inner rings indicate that, although the source material in both regions is mainly water ice, the makeup of the plasma depends critically on ion-molecule chemistry and transport. Thus, the inner plasmasphere, which is rich in water-group ions including H_3O^+, is somewhat analogous to a cometary coma, and the rings' O_2 atmosphere is similar in composition to those of Europa and Ganymede. The presence of low-energy N^+ deep in the magnetosphere points to local sources of nitrogen. The implications of our findings for the cycle of neutral gas and plasma generation and their redistribution are not clear and must await further measurements and modeling. Future orbits (there are 74 remaining) will allow us to substantiate and improve the accuracy of these results.

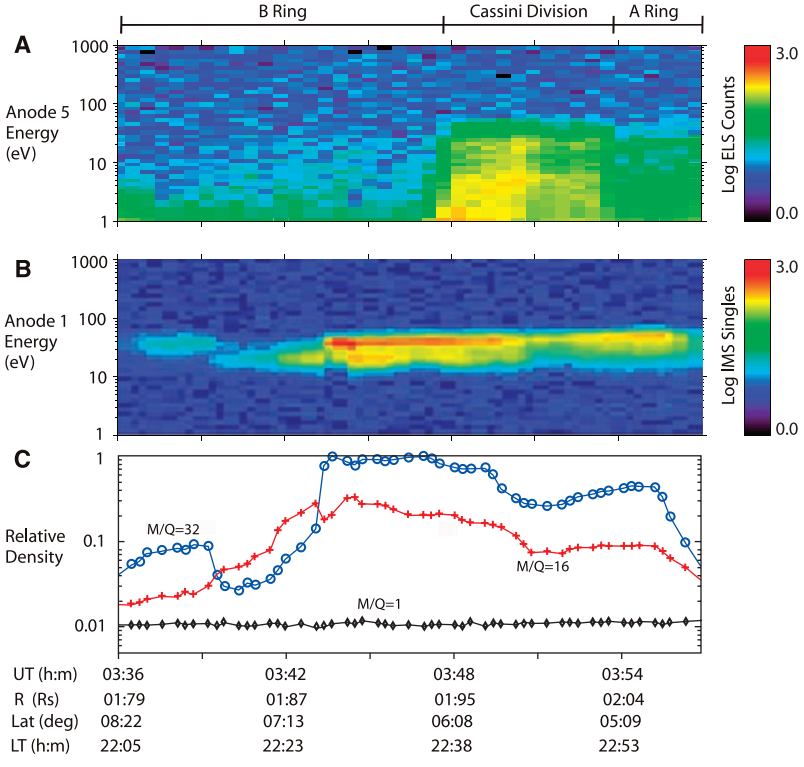


Fig. 4. Color spectrograms and relative ion densities over the A and B rings. (**A**) Electron energy-time spectrogram in the same format as Fig. 1. (**B**) Ion energy-time spectrogram in the same format as Fig. 1. (**C**) Ion density normalized to the maximum density and plotted in arbitrary units. The density was determined from the total IMS ion flux and the assumption of ion corotation. The identification of *M*/*Q* as 16 and 32 ions is consistent with this assumption. Because of the limited IMS high-voltage range during this period, true mass-spectral identification from TOF data is not yet available. The Cassini ring labels correspond to the mapping of magnetic field lines between the spacecraft and the ring plane. At 03:57 UT, the spacecraft began a turn that prevented CAPS from viewing in the direction needed to detect corotating ions. Valid electron measurements extending past 03:57 UT show that the ionosphere extends out to the edge of the A ring.

References and Notes

1. L. A. Frank, B. G. Burek, K. L. Ackerson, K. L. Wolfe, J. D. Mihalov, *J. Geophys. Res.* **85**, 5695 (1980).
2. H. S. Bridge *et al.*, *Science* **212**, 217 (1981).
3. H. S. Bridge *et al.*, *Science* **215**, 563 (1982).
4. E. C. Sittler, K. W. Ogilvie, J. D. Scudder, *J. Geophys. Res.* **88**, 8847 (1983).
5. A. J. Lazarus, R. L. McNutt, *J. Geophys. Res.* **88**, 8831 (1983).
6. J. D. Richardson, E. C. Sittler, *J. Geophys. Res.* **95**, 12019 (1990).
7. S. Maurice *et al.*, *J. Geophys. Res.* **101**, 15211 (1996).
8. J. D. Richardson, *Rev. Geophys.* **36**, 501 (1998).
9. D. T. Young *et al.*, *Space Sci. Rev.* **114**, 1 (2004).
10. The IMS, ELS, and IBS measure particle direction-of-arrival over one hemisphere with resolutions of 8° × 20°, 5° × 20°, and 1.5° × 1.5°, respectively. Two-dimensional energy-angle spectra are measured continuously every 2 s by the ELS, and every 4 s by the IMS. IBS energy spectra are measured every 2 s. Electron distributions are often isotropic, or nearly so, during the period reported here and can be observed by the ELS regardless of spacecraft orientation, which can be highly variable. Ion flux, on the other hand, is often observed to be directional, and our ability to view it depends on spacecraft orientation.
11. Relative to Voyager measurements, CAPS has greatly improved capabilities in three critical areas. (i) The IMS is a true mass spectrometer, not generally dependent on assumptions about ion *E*/*Q* distributions for derivation of ion *M*/*Q*. (ii) Voyager gaps in *E*/*Q* coverage <10 V and 6 to ~30 kV are filled. (iii) Temporal coverage is improved from one 3.8-s energy spectrum taken every 96 s to contiguous spectra taken every 2 s (by the ELS and IBS) or 4 s (by the IMS). One other important difference is that CAPS observations extend to as close as 1.7 R_S over the rings compared with 2.7 R_S for Voyager 2.
12. H. T. Smith, R. E. Johnson, V. I. Shematovich, *Geophys. Res. Lett.* **31**, L16804 (2004).
13. A corotating plasma with the same angular velocity as the planet has an ion energy of $0.5M(R/R_S)^2$ eV, where *M* is ion mass in amu. This formula is used to calculate the corotation curves superimposed on Figs. 1 and 2.
14. A. Nishida, *J. Geophys. Res.* **71**, 5669 (1966).
15. T. W. Hill, A. J. Dessler, F. C. Michel, *Geophys. Res. Lett.* **1**, 3 (1974).
16. J. D. Richardson, *Geophys. Res. Lett.* **22**, 1177 (1995).
17. J. D. Richardson, A. Eviatar, M. A. McGrath, V. M. Vasyliunas, *J. Geophys. Res.* **103**, 20245 (1998).
18. B. H. Mauk, C. I. Meng, *J. Geophys. Res.* **88**, 10011 (1983).
19. B. H. Mauk, D. J. Williams, R. W. McEntire, *Geophys. Res. Lett.* **24**, 2949 (1997).
20. D. H. Pontius Jr., T. W. Hill, M. E. Rassbach, *Geophys. Res. Lett.* **13**, 1097 (1986).
21. M. Dougherty *et al.*, *Science* **307**, 1266 (2005).
22. W. D. Cummings, A. J. Dessler, *J. Geophys. Res.* **72**, 1007 (1967).
23. D. J. Knudsen, M. C. Kelley, J. F. Vickrey, *J. Geophys. Res.* **97**, 77 (1992).
24. A. Hasegawa, K. Mima, *J. Geophys. Res.* **83**, 1117 (1978).
25. A. Eviatar, *J. Geophys. Res.* **89**, 3821 (1984).
26. M. L. Delitsky, A. L. Lane, *J. Geophys. Res.* **107**, 5093 (2002).
27. J. S. Kargel, S. Pozio, *Icarus* **119**, 385 (1996).

28. S. Jurac *et al.*, *Geophys. Res. Lett.* **29**, 2172 (2002).
29. R. E. Johnson *et al.*, *Icarus* **77**, 311 (1989).
30. D. E. Shemansky, P. Matheson, D. T. Hall, H.-Y. Hu, T. M. Tripp, *Nature* **363**, 329 (1993).
31. H. Balsiger *et al.*, *Nature* **321**, 330 (1986).
32. S. Jurac, R. E. Johnson, J. D. Richardson, C. Paranicas, *Planet. Space Sci.* **49**, 319 (2001).
33. D. A. Gurnett *et al.*, *Science* **307**, 1255 (2005).
34. L. W. Esposito *et al.*, in *Saturn*, T. Gehrels, M. S. Edwards, Eds. (Univ. of Arizona Press, Tucson, AZ, 1984), pp. 463–545.
35. This identification was made without time-of-flight (TOF) spectra because at this time, the IMS TOF system was not operating at full voltage and so was out of calibration. Further analysis of the spectra will allow us to positively identify the ion species.
36. W.-I. Ip, *Icarus* **115**, 295 (1995).
37. R. E. Johnson, T. I. Quickenden, P. D. Cooper, A. J. McKinley, C. Freeman, *Astrobiology* **3**, 823 (2003).
38. R. E. Johnson *et al.*, in *Jupiter: The Planet, Satellites, and Magnetosphere*, F. Bagenal, T. Dowling, W. B. McKinnon, Eds. (Cambridge Univ. Press, Cambridge, 2004), pp. 485–512.
39. M. Shi *et al.*, *J. Geophys. Res.* **100**, 26387 (1995).
40. R. E. Johnson, T. I. Quickenden, *J. Geophys. Res.* **102**, 10985 (1997).
41. The CAPS instrument and investigation represent the culmination of 16 years of effort by a team of 170 engineers and scientists at 15 institutions in 6 countries. We wish to thank all team members for their contributions and dedication to the CAPS and Cassini effort. Data reduction and analysis in the United States is supported by NASA/Jet Propulsion Laboratory under contract 1243218 with the Southwest Research Institute. Work at Los Alamos was performed under the auspices of the U.S. Department of Energy. We also wish to thank national funding agencies in Finland, France, Hungary, Norway, and the UK Particle Physics and Astronomy Research Council for their support.

7 October 2004; accepted 10 December 2004
10.1126/science.1106151

REPORT

Cassini Magnetometer Observations During Saturn Orbit Insertion

M. K. Dougherty,[1]* N. Achilleos,[1] N. Andre,[2] C. S. Arridge,[1] A. Balogh,[1] C. Bertucci,[1] M. E. Burton,[3] S. W. H. Cowley,[4] G. Erdos,[5] G. Giampieri,[1] K.-H. Glassmeier,[6] K. K. Khurana,[7] J. Leisner,[7] F. M. Neubauer,[8] C. T. Russell,[7] E. J. Smith,[3] D. J. Southwood,[9] B. T. Tsurutani[3]

Cassini's successful orbit insertion has provided the first examination of Saturn's magnetosphere in 23 years, revealing a dynamic plasma and magnetic environment on short and long time scales. There has been no noticeable change in the internal magnetic field, either in its strength or its near-alignment with the rotation axis. However, the external magnetic field is different compared with past spacecraft observations. The current sheet within the magnetosphere is thinner and more extended, and we observed small diamagnetic cavities and ion cyclotron waves of types that were not reported before.

The first in situ observations from Saturn's magnetosphere in 23 years were obtained during Cassini's Saturn orbit insertion on 30 June 2004. The magnetometer instrument (*1*) (MAG) obtained data on upstream waves, bow shock, magnetosheath, magnetopause, magnetospheric currents, and waves, as well as the planetary magnetic field. These data are consistent with measurements made on earlier missions, provide more detail of some parameters such as the planetary magnetic field, including its possible secular variation, and reveal some features in the external field not previously reported.

The solar wind controls the size of the magnetosphere and the dynamics of its outer reaches. Because the solar wind speed is supersonic, the deflection of the solar wind occurs via a standing bow shock that compresses and heats the solar wind, forming the magnetosheath. The inner edge of the magnetosheath, the magnetopause, marks the outer boundary of the region controlled by the planetary magnetic field. Saturn's bow shock is of intrinsic interest because it is expected to be much stronger than that of Earth. The magnetopause is important because it controls the coupling of the solar wind flow to the magnetosphere, principally, we expect, through the process known as reconnection (*2*). The locations of both boundaries are determined by the dynamic pressure of the solar wind and the combined plasma and magnetic pressure of the magnetosphere.

The boundaries were observed to be very dynamic. We measured a total of 17 bow shock and 7 magnetopause crossings (Fig. 1) on the inbound and outbound passages. Bow shock crossings were identified by abrupt increases in the magnetic field magnitude where the solar wind was compressed and decelerated. As Cassini approached Saturn near 08:00 local time (LT), it crossed the bow shock on seven separate occasions, starting at 09:45 universal time (UT) on 27 June 2004 at a distance of $49.15R_S$ [1 R_S = 60,268 km (*3*)]. The last inbound bow shock crossing occurred at 05:38:30 UT on 28 June 2004 at a distance of $40.5R_S$. At least 10 crossings of the bow shock were observed on the outbound leg in the magnetometer data, with the earliest bow shock crossing observed on 7 July 2004 at a radial distance range of $56R_S$ and the final bow shock crossing occurring on 14 July 2004 at $85R_S$. Saturn's bow shock is expected to be a strong quasi-perpendicular shock. A large overshoot in the magnetic field magnitude is a consistent feature of the shock crossings and is a typical feature of supercritical planetary bow shocks (*4*). These observations afford us the opportunity to study bow shocks at high Mach numbers, which are rarely observed on Earth. Indeed, the Mach number can be inferred from the amplitude of the overshoot $\delta B/B_0$ (*5*). The average value for the bow shocks observed is ~1.5, consistent with a fast magnetosonic Mach number as high as 8.

On the Saturnian side of the bow shock lies the magnetosheath, where the magnetic field is very turbulent and dominated by wave activity. The entrance into the magnetosphere was preceded by multiple crossings of the magnetopause on 28 June and 29 June 2004. On the outbound pass, the spacecraft exited the magnetosphere near 05:00 LT. The first excursions into the inner magnetosheath were characterized by the presence of mirror mode structures. The multiple crossings of these plasma boundaries reveal the dynamic character of Saturn's outer magnetosphere in response to variations in the solar wind ram pressure.

Before the Cassini arrival at Saturn, there had been three previous flybys: Pioneer 11 (*6*) and Voyager 1 and 2 (*7*, *8*). The spatial coverage of these flybys was limited, with the inbound trajectories all being in the noon sector and outbound Pioneer 11 and Voyager 1

[1]Blackett Laboratory, Imperial College London, SW7 2AZ, UK. [2]CESR, 31028 Toulouse, France. [3]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA. [4]Department of Physics and Astronomy, Leicester University, Leicester, LE1 7RH, UK. [5]KFKI Research Institute for Particle and Nuclear Physics, H-1525 Budapest, Hungary. [6]Technische Universitat Braunschweig, D-38106 Braunschweig, Germany. [7]University of California Los Angeles Institute of Geophysics and Planetary Physics, Los Angeles, CA 90024, USA. [8]Institute for Geophysics and Meterology, Köln University, 50923 Köln, Germany. [9]European Space Agency, 75738 Paris, France.

*To whom correspondence should be addressed. E-mail: m.dougherty@imperial.ac.uk

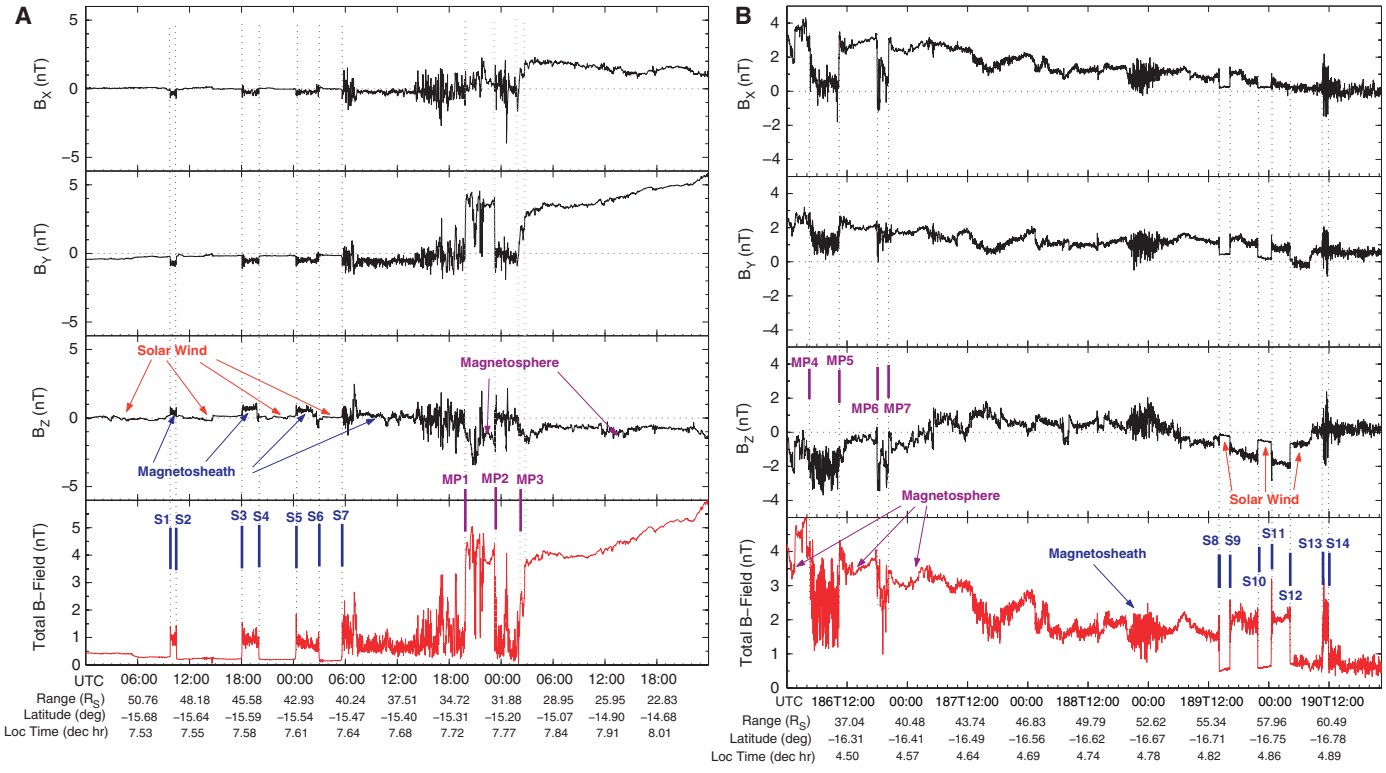


Fig. 1. (**A**) The magnetic field (*B*) between 27 and 29 June 2004 as the spacecraft approached Saturn. The top three sections show the three components of the field, with the magnitude depicted in the bottom section in KSM coordinates (*24*). Multiple bow shock (S1, S2, ...) and magnetopause (MP1, MP2, ...) crossings are indicated by vertical lines and labels. Information detailing the radial distance, latitude, and local time of the spacecraft with respect to Saturn are given. Time is given in universal time coordinated (UTC). (**B**) The magnetic field between 4 and 8 July 2004 on the outbound passage after Saturn orbit insertion.

exiting in the dawn meridian (Fig. 2). The inbound Cassini trajectory occurred at an earlier local time, and its outbound passage was similar to that of Voyager 2, exiting at high southern latitudes. A measure of the variability in the position of the different plasma boundaries can be assessed by comparing the locations of the Cassini crossings with their nominal position as deduced from the previous observations (*9*, *10*). The Cassini inbound crossings occurred further from Saturn than their average expected locations, revealing an expanded magnetosphere at the time of the inbound encounter (Fig. 2) with the first magnetopause crossing occurring a day earlier than expected. The state of the magnetosphere during the outbound part of the orbit was very different. The magnetopause and bow shock crossings were located inside their respective average surfaces, in agreement with a magnetosphere in a compressed regime. This regime is consistent with predictions (*11*) from upstream magnetic observations during the approach to Saturn. The entry into the magnetosphere was identified from the draping of the magnetic field, which is compatible with a dipole. In addition, an expansion of the magnetosphere occurred during the outbound passage of Cassini, with the magnetopause moving back out over the spacecraft and with the last bow shock occurring close to the expected nominal distance at $85R_S$.

In situ measurements of the internal magnetic field at Saturn have been obtained from the three previous flyby missions. The Cassini mission as an orbiter will provide a much more complete three-dimensional sampling of the magnetic field over the next 4 years. Until more complete measurements of the Saturnian magnetosphere are obtained during future orbits, the insertion observations (Fig. 3) can essentially be treated as a single flyby. In particular, for the inversion of the internal field we used only data within $8R_S$ and assumed an axisymmetric field configuration up to degree 3. External terms were also included by adding an axisymmetric ring current whose presence was revealed by analysis of previous data. Previous analyses led to the SPV and Z3 models (*12*, *13*), which included terms corresponding to a uniform external field in a source-free region. However, we have found that such a uniform field is not adequate for describing the field in and near the ring current, and therefore we have used the more realistic Connerney model adapted to Saturn (*14*, *15*).

Precise knowledge of the gain and pointing of the magnetometer is important for the accurate inversion of the measurements to obtain the internal magnetic field. We performed extensive ground calibrations before launch, checked the calibrations during the Earth flyby, and monitored the calibrations since launch with the magnetic field produced by coils both remote from and near the sensors. We estimate the uncertainty is presently 0.1% in gain and 0.1° in pointing. The magnetic field data were analyzed by using a standard singular value decomposition of the inversion matrix. Only zonal terms up to degree 3 are included in addition to the four disk parameters (namely, total current, inner and outer radii, and thickness). The internal field coefficients are shown (Table 1) alongside those from past models.

The quantity $g_{20}/(2g_{10})$, where g_{20} is the quadrupole moment and g_{10} the dipole moment, can be interpreted as a northward displacement of the dipole by $0.037R_S$. The comparable results for the SPV, Z3, and GD (*15*) models are $0.037R_S$, $0.038R_S$, and $0.037R_S$, respectively. The implication is that the magnetic and rotational equators do not coincide. However, the relatively large octupole moment cannot be explained by this offset. After accounting for the uncertainties associated with the estimated parameters, we were not able to determine any secular change in the internal field zonal coefficients. Also, we noticed that the azimuthal field component was barely visible near closest approach, which confirms the near-axial symmetry of the field. Although we have not considered the azimuthal component in our analysis because of the limited spatial sampling of the insertion trajectory, a previous study (*15*) has revealed the likely presence of a small azimuthal field

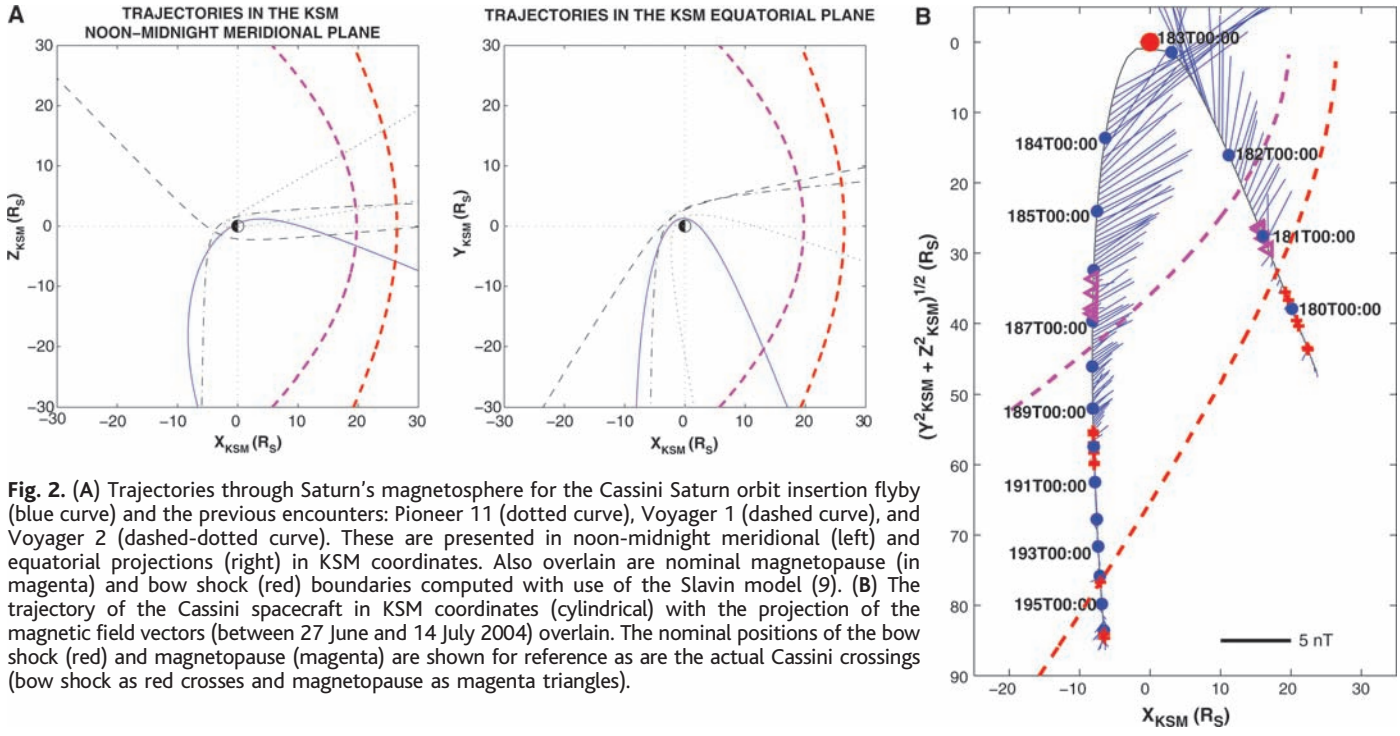


Fig. 2. (**A**) Trajectories through Saturn's magnetosphere for the Cassini Saturn orbit insertion flyby (blue curve) and the previous encounters: Pioneer 11 (dotted curve), Voyager 1 (dashed curve), and Voyager 2 (dashed-dotted curve). These are presented in noon-midnight meridional (left) and equatorial projections (right) in KSM coordinates. Also overlain are nominal magnetopause (in magenta) and bow shock (red) boundaries computed with use of the Slavin model (*9*). (**B**) The trajectory of the Cassini spacecraft in KSM coordinates (cylindrical) with the projection of the magnetic field vectors (between 27 June and 14 July 2004) overlain. The nominal positions of the bow shock (red) and magnetopause (magenta) are shown for reference as are the actual Cassini crossings (bow shock as red crosses and magnetopause as magenta triangles).

of internal origin, in accordance with theoretical constraints.

In the inner magnetosphere, the largest contribution to the external field was from the ring current, whose magnitude was as large as ~20 nT at the point where Cassini crossed the inner boundary of the disk ($r \sim 6R_S$). The estimated inner radius and current density of the disk are in very good agreement with our previous estimates; however, there are indications that the disk is thinner (half-thickness < $2R_S$) as well as more elongated (outer radius > $20R_S$) than in the past.

As in the case of the Earth and Jupiter, the magnetic field at Saturn is generated by three different current systems: in the core, in the magnetospheric plasma, and from the solar wind interaction with the planetary magnetic field. Subtracting the planetary magnetic field model allows us to examine the contribution of the various currents flowing inside the magnetosphere and on its boundaries. This differenced field ($\Delta\mathbf{B} = \mathbf{B}_{obs} - \mathbf{B}_{int}$) shown at radial distances greater than $4R_S$ can be seen in Fig. 4 in the black trace. Overlain on this figure are two different models of the external field in Saturn's magnetosphere. The blue trace is the best fit Connerney disk as described above, and the red trace is a new global magnetospheric model that models the external field, i.e., considering the internal field to be a fixed, known quantity.

In this global model, we retain the Connerney current disk to model the equatorial azimuthal current and used parameters obtained from the Voyager 1 flyby. The model also includes fields to represent the effect of magnetopause currents, which are modeled by the source-surface method (*16*) used in the Tsyganenko models of the terrestrial magnetosphere (*17*–*19*). We shield the magnetic field so that it is tangential to the magnetopause boundary at all points over that surface, which is represented by a prolate hemi-ellipsoid (*20*) for the nose region and a cylinder for the tail. The field due to the shielding of the dipole is represented with use of cylindrical harmonics and that due to the shielding of the disk with use of Cartesian harmonics. These representations allow the model to have an analytical dependence on the tilt angle between the dipole axis and the solar wind flow and on the solar wind dynamic pressure. The magnetopause stand-off distance, a function of the dynamic pressure, was set to be consistent with the location of the final inbound magnetopause crossing.

The global model does not fit the observed data well, and the current disk model suggests that the current disk field has changed compared with that of the Voyager 1 epoch. The magnitude of the azimuthal field in the outer magnetosphere caused by magnetopause currents is larger than that estimated from the model. This has also been seen in the previous flyby data sets (*21*) and may hint at further dynamical processes. At present, the model does not include a magnetotail current, which means we are underestimating the field in the tail region. Neither model accurately accounts for both the inbound and outbound observations simultaneously. This feature of the Saturnian magnetosphere has been noted in modeling studies and indicates an equatorial current disk that has local time structure and asymmetry. However, the compression of the magnetosphere during the Cassini flyby probably also altered the equatorial current distribution.

When Cassini was immersed within the plasma sheet, changes in the magnetic field

Table 1. Zonal harmonics from Saturn orbit insertion compared with past models (*25*). All coefficients are in reference to R_S = 60, 268 km. The first three models are axially symmetric up to degree 3; the fourth one includes also nonaxial terms up to the same degree.

Multipole term	Cassini	SPV	Z3	GD
g_{10}	21,084	21,225	21,248	21,232
g_{20}	1544	1566	1613	1563
g_{30}	2150	2332	2683	2821

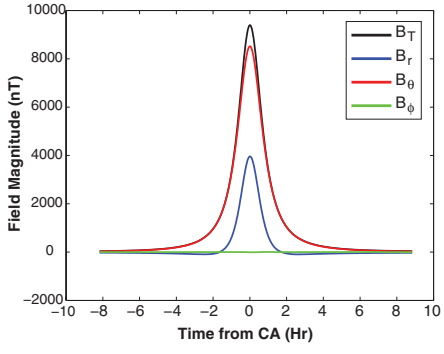


Fig. 3. The fluxgate magnetometer data from the closest approach (CA) period, with the three components and magnetic field magnitude shown.

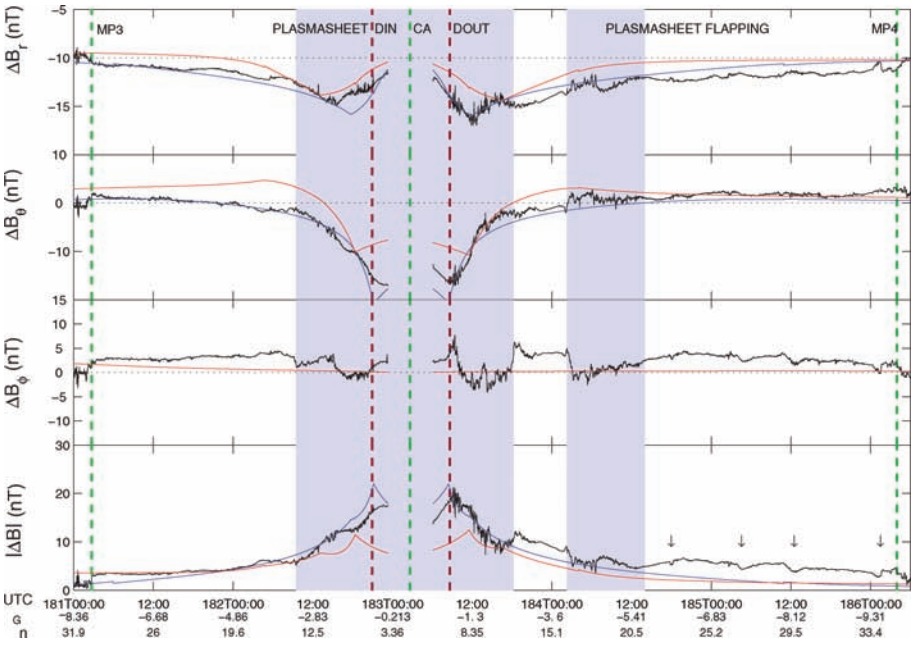


Fig. 4. Differenced field (black) in spherical polar coordinates obtained by subtracting the planetary field, as derived by the inversion of the Saturn orbit insertion data. This field represents the effect of magnetospheric current systems, and two models of this external field are overlaid. In blue is the best fit Connerney current disc (*14*) to the data, and in red is a new global model of Saturn's magnetospheric field. The last magnetopause crossing inbound and the first outbound crossing and perikron are indicated by vertical green dashed lines. The vertical crimson lines marked DIN and DOUT are crossings of Dione's L-shell inbound and outbound, respectively. The arrows indicate a quasi-periodic signature, which may also be plasma sheet–related. Shaded regions indicate where Cassini was immersed in the plasma sheet. Distance from the equatorial plane (z_{KG}) and range (r) are given in units of Saturn radii.

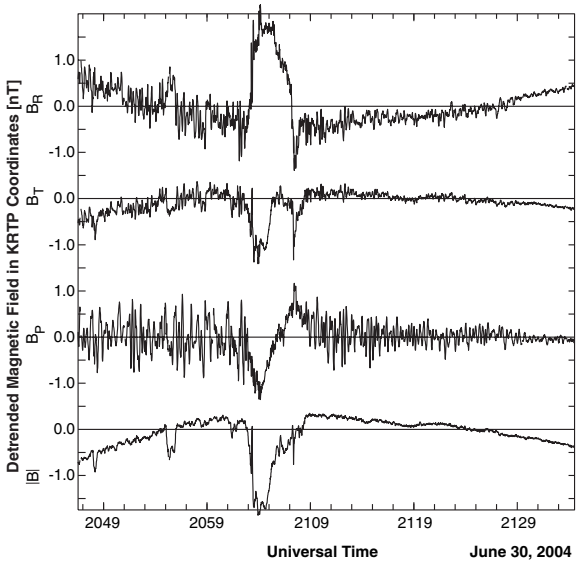


Fig. 5. Magnetic field strength through the largest and last diamagnetic depression seen on the inbound trajectory. Data have been detrended and is in KRTP, spherical polar Saturn centered coordinates, when R is radial array from Saturn, P is azimuthal, and T is meridional.

configuration occurred with an enhanced level of small-scale magnetic fluctuations. Outside of these regions, the Cassini spacecraft was located in the outer higher-latitude magnetosphere, a region with less plasma that is similar to the lobe regions of the Earth's magnetosphere. The plasma sheet crossings occurred at planetocentric distances of $15R_S$ inbound and $12R_S$ outbound and are associated with sharp transitions in the azimuthal field. The observed field fluctuations reveal a dynamic region associated with plasma sources, transport processes, and related wave generation. Closer to the planet the field is more dominated by the internal source. Outbound from about $16R_S$ and coincident with the Titan far encounter, the spacecraft may have again crossed the plasma sheet at high latitude, possibly caused by a dynamic solar wind–driven motion of the sheet. The field at these latitudes and distances is also tail-like, with a more stretched-out configuration.

On both the inbound and outbound legs of the pass, near but inside the orbit of Dione at $6.27R_S$, MAG encountered diamagnetic depressions in which some of the magnetic pressure was replaced by increased plasma pressure. On each leg the innermost depression was the largest. Such diamagnetic cavities were not reported for the Voyager and Pioneer flybys.

The largest of the depressions occurred at 21:05 UT on 30 June 2004 at a radial distance of $5.93R_S$, a local time of 09:45 LT, and a latitude of –9.5° (Fig. 5). Dione was at 15:08 LT. The maximum depth of the magnetic depression in the 95-nT background field was produced by a plasma with an energy density of 500 eV cm^{-3}. If produced by 1 keV electrons, a plasma density of about 0.5 electrons cm^{-3} would be required. On the outbound leg at 08:02 UT, $5.5R_S$, 01:53 LT, and –9° latitude, a similar depression was seen in a 115-nT background field corresponding to a plasma energy density of 300 eV cm^{-3}. If this is the outer boundary of a cold plasma torus, we would expect that the source of the cold ions would be the E ring and the hot electrons outside the cold torus region would be produced close to the ionosphere, where the field-aligned current system needed to enforce corotation of the plasma closes. Surrounding the two largest depressed field regions were relatively strong waves polarized in the direction transverse to the magnetic field. There was essentially no compressional component above the noise level of the instrument. Pioneer 11 (*22*) also saw transverse waves but beyond the orbit of Dione. The transverse Cassini waves are accompanied by a narrow band of left-hand cyclotron waves propagating very close to the direction of the magnetic field at the H_2O^+ gyro-frequency. Although seen on both sides of the depressions and on both legs, this narrow band of ion cyclotron waves makes very little contribution to the total wave power. The water group ion cyclotron waves are presumably associated with ion pickup from the E ring, but we do not have an obvious source for the strong linear transverse waves that surround the cavity. These are associated with the cavities because their amplitudes decrease with distance from the cavities.

At greater distances from the planet, the ratio of depressed to enhanced flux tubes increased until the field strength appeared to be continually depressed. The phenomenon in the Jovian magnetosphere that most nearly resembles these diamagnetic depressions is at the outer edge of the Io torus, at the boundary between the cold dense torus plasma and hotter magnetospheric plasma where flux tube interchange appears to take place (*23*). Alternatively this region may be where corotation of the magnetospheric plasma can no longer

be maintained by the ionosphere. The resulting slippage in the ionosphere heats the ionospheric electrons, producing high beta conditions along the entire flux tube.

Quite distinct from the waves surrounding the cavities was a 100-min-long burst of ion cyclotron waves (from 06:10 to 07:50 UT on 30 June 2004) unlike any seen on Voyager and Pioneer that appeared 5 hours after the Cassini engine stopped firing. These waves were limited to the frequency band expected for the singly ionized products of the engine exhaust H_2O, N_2, CO, and CO_2. We do not expect these waves to be present (except for H_2O^+ that may be associated with the icy satellites' environments) in the natural plasma. We note that Cassini's engine deposited over 850 kg of fuel in the Saturnian magnetosphere and, as it ionized and traversed the magnetosphere, it would produce a cloud of ions with energy predominantly transverse to the magnetic field, that is, a ring beam. The energy of these pickup ions at the locations where the waves were seen is comparable to that of the pickup ions in the Io torus. This may be the first detection of artificially induced plasma waves in a magnetosphere other than that of Earth.

References and Notes

1. M. K. Dougherty *et al.*, *Space Sci. Rev.* **114**, 331 (2004).
2. J. W. Dungey, *Phys. Rev. Lett.* **6**, 47 (1961).
3. International Astronomical Union standard definition of the Saturn equatorial body radius.
4. C. T. Russell, M. M. Hoppe, W. A. Livesey, *Nature* **296**, 45 (1982).
5. The ratio $\delta B/B0$ is the enhancement of the overshoot magnetic field normalized by the magnetosheath field strength away from the overshoot.
6. E. J. Smith *et al.*, *J. Geophys. Res.* **85**, 5655 (1980).
7. N. F. Ness *et al.*, *Science* **212**, 211 (1981).
8. N. F. Ness *et al.*, *Science* **215**, 558 (1982).
9. J. A. Slavin, E. J. Smith, P. Gazis, J. Mihalov, *Geophys. Res. Lett.* **10**, 9 (1983).
10. J. A. Slavin, E. J. Smith, J. Spreiter, S. Stahara, *J. Geophys. Res.* **90**, 6275 (1985).
11. C. M. Jackman *et al.*, *J. Geophys. Res.* **109**, A11203 (2004).
12. L. Davis, E. J. Smith, *J. Geophys. Res.* **95**, 15257 (1990).
13. M. H. Acuna, J. E. P. Connerney, N. F. Ness, *J. Geophys. Res.* **88**, 8771 (1983).
14. J. E. P. Connerney, M. H. Acuna, N. F. Ness, *J. Geophys. Res.* **88**, 8779 (1983).
15. G. Giampieri, M. K. Dougherty, *Geophys. Res. Lett.* **31**, 16701 (2004).
16. M. Schultz, M. C. McNab, *Geophys. Res. Lett.* **14**, 182 (1987).
17. N. A. Tsyganenko, *Planet. Space Sci.* **37**, 5 (1989).
18. N. A. Tsyganenko, *J. Geophys. Res.* **100**, 5599 (1995).
19. N. A. Tsyganenko, *J. Geophys. Res.* **107**, SMP12-1 (2002).
20. The dayside of the magnetopause is approximately ellipsoidal, but this is not the case in the tail, which is nearly circular in cross section. To accurately approximate the shape of the magnetopause, the tailward half of an ellipsoid is replaced with a cylinder; thus, the boundary is a composite surface composed of a hemi-ellipsoid and a cylinder. The join between these two surfaces is at the semi-minor axis of the prolate ellipsoid (nose portion).
21. E. J. Bunce, S. W. H. Cowley, J. A. Wild, *Ann. Geophys.* **21**, 1709 (2003).
22. E. J. Smith, B. T. Tsurutani, *J. Geophys. Res.* **88**, 7831 (1983).
23. C. T. Russell *et al.*, *Adv. Space Res.* **26**, 1489 (2000).
24. The KSM (Kronocentric Solar Magnetospheric) coordinate system has Saturn at the origin, with the $+X$ axis directed toward the Sun; Z defined such that Saturn's rotation and magnetic axis lies in the XZ plane (with $+Z$ pointing close to northward), and Y lying in Saturn's rotational and magnetic equatorial plane.
25. The reference for the SPV model is (*13*); for the Z3 model, (*14*); and the GD model, (*15*).
26. We wish to acknowledge the Cassini Project and operations team from the Jet Propulsion Laboratory, as well as the many people at our home institutions for their design, engineering, and software support over many years on behalf of this investigation. The contributions to the MAG team have been supported by the Particle Physics and Astronomy Research Council in the UK, Deutsches Zentrum für Luft-und Raumfahrt in Germany, and NASA in the U.S.

6 October 2004; accepted 7 December 2004
10.1126/science.1106098

REPORT

Dynamics of Saturn's Magnetosphere from MIMI During Cassini's Orbital Insertion

S. M. Krimigis,[1]* D. G. Mitchell,[1] D. C. Hamilton,[2] N. Krupp,[3] S. Livi,[1] E. C. Roelof,[1] J. Dandouras,[4] T. P. Armstrong,[5] B. H. Mauk,[1] C. Paranicas,[1] P. C. Brandt,[1] S. Bolton,[6] A. F. Cheng,[1] T. Choo,[1] G. Gloeckler,[2] J. Hayes,[1] K. C. Hsieh,[7] W.-H. Ip,[8] S. Jaskulek,[1] E. P. Keath,[1] E. Kirsch,[3] M. Kusterer,[1] A. Lagg,[3] L. J. Lanzerotti,[9,10] D. LaVallee,[1] J. Manweiler,[5] R. W. McEntire,[1] W. Rasmuss,[5] J. Saur,[1] F. S. Turner,[1] D. J. Williams,[1] J. Woch[3]

The Magnetospheric Imaging Instrument (MIMI) onboard the Cassini spacecraft observed the saturnian magnetosphere from January 2004 until Saturn orbit insertion (SOI) on 1 July 2004. The MIMI sensors observed frequent energetic particle activity in interplanetary space for several months before SOI. When the imaging sensor was switched to its energetic neutral atom (ENA) operating mode on 20 February 2004, at $\sim 10^3$ times Saturn's radius R_S (0.43 astronomical units), a weak but persistent signal was observed from the magnetosphere. About 10 days before SOI, the magnetosphere exhibited a day-night asymmetry that varied with an ~11-hour periodicity. Once Cassini entered the magnetosphere, in situ measurements showed high concentrations of H^+, H_2^+, O^+, OH^+, and H_2O^+ and low concentrations of N^+. The radial dependence of ion intensity profiles implies neutral gas densities sufficient to produce high loss rates of trapped ions from the middle and inner magnetosphere. ENA imaging has revealed a radiation belt that resides inward of the D ring and is probably the result of double charge exchange between the main radiation belt and the upper layers of Saturn's exosphere.

The magnetosphere of Saturn was discovered by Pioneer 11 in 1979 (*1*, *2*) and was investigated in detail by Voyager 1 in 1980 (*3*) and Voyager 2 in 1981 (*4*). The primary science objectives of MIMI (*5*) are to determine the global configuration and dynamics of hot plasma in the magnetosphere of Saturn through energetic neutral particle imaging of the ring current, radiation belts, and neutral clouds; to study the sources of plasmas and energetic ions through in situ measurements of energetic ion composition, spectra, charge state, and angular distribution; to search for, monitor, and analyze magnetospheric spatial and time variations; to determine through imaging and composition studies the magnetosphere-satellite interactions at Saturn; to understand the formation of clouds of neutral hydrogen, nitrogen, and water products; and to study Titan's comet-like interaction with Saturn's magnetosphere and the solar wind. The MIMI instrument (*5*) comprises three sensors that measure particles in specific energy ranges: (i) the Ion and Neutral Camera (INCA), which measures ion and

[1]Applied Physics Laboratory, Johns Hopkins University, Laurel, MD 20723, USA. [2]Department of Physics, University of Maryland, College Park, MD 20742, USA. [3]Max-Planck-Institut für Sonnensystemforschung, D-37191 Lindau, Germany. [4]Centre D'Etude Spatiale Des Rayonnements, F-31028 Toulouse, France. [5]Fundamental Technologies Inc., Lawrence, KS 66049, USA. [6]Jet Propulsion Laboratory, Pasadena, CA 91109, USA. [7]Department of Physics, University of Arizona, Tucson, AZ 85721, USA. [8]Graduate Institute of Astronomy, National Central University, Jhongli City, Republic of China. [9]Bell Laboratories, Murray Hill, NJ 07974, USA. [10]New Jersey Institute of Technology, Newark, NJ 07102, USA.

*To whom correspondence should be addressed. E-mail: tom.krimigis@jhuapl.edu

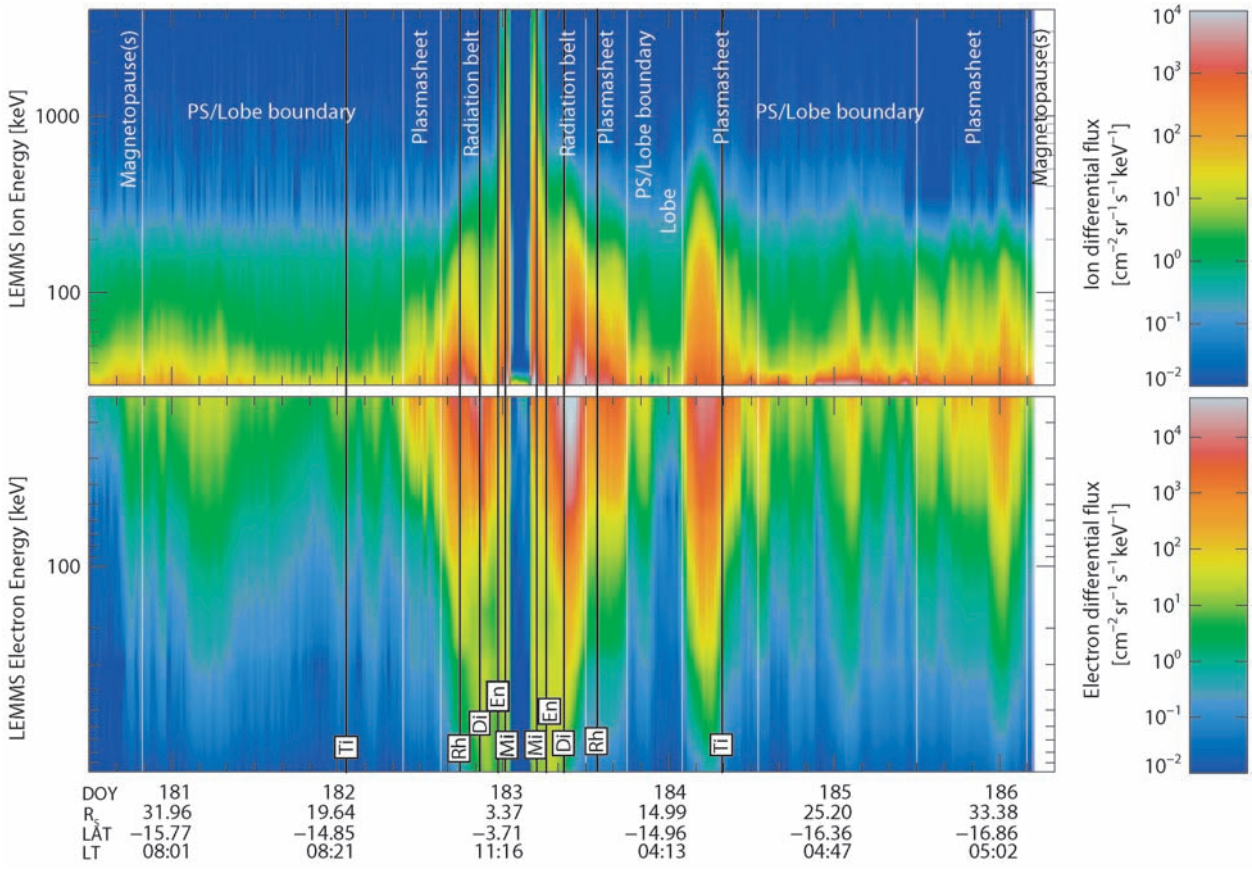


Fig. 1. Dynamic spectrograms (energy versus time) of LEMMS ion and electron intensities. Electron energy is plotted increasing downward for comparison with the features in the ion population. Magnetospheric regions are labeled and defined in the text. The general particle morphology is that of a well-defined radiation belt outside the ABC rings, transitioning abruptly to an extended, highly dynamic plasma sheet that extends outward to the magnetopause crossings (both inbound and outbound). Notations for moons: Ti, Titan; Rh, Rhea; Di, Dione; En, Enceladus; Mi, Mimas.

neutral species (~3 to 200 keV per nucleon); (ii) the Charge Energy Mass Spectrometer (CHEMS), which measures ions and their charge states (3 to 230 keV per charge); and (iii) the Low Energy Magnetospheric Measurement System (LEMMS), which measures ions (0.02 to 18 MeV) and electrons (0.015 to 1 MeV). LEMMS also measures high-energy electrons ($E > 3$ MeV) and protons ($1.6 < E <$ 160 MeV) from the back end of the dual field-of-view telescope.

The LEMMS ion and electron intensities (Fig. 1) measured by MIMI in several magnetospheric regions during Cassini's inbound and outbound SOI pass (29 June to 4 July 2004) are similar to intensities measured by the Voyager 1 and 2 spacecraft (*3*, *4*). Measurements of the intensities and the magnetic field (*6*) allow us to define three different regions of the magnetosphere: the radiation belt, the plasma sheet, and the lobe. In general, the ions and electrons exhibit the same type of behavior. Essentially no energetic particles are detected above the rings, because they are effectively absorbed by solid material.

The radiation belt consists of energetic particles (tens to thousands of keV) with intensities that are locally modulated via absorption by saturnian satellites. Otherwise, ion and electron intensities generally exhibit smooth histories, although there are identifiable nondispersive intensity increases and decreases. The magnetic field strength and direction are consistent with those of a dipole (*6*). The highest ion and electron intensities (outside of the inner magnetosphere) are confined to the near-equatorial plasma sheet. Relatively abrupt increases in intensity occur as Cassini enters the plasma sheet region at the beginning of the plasma sheet intervals. Conversely, relatively abrupt decreases in intensity occur as Cassini leaves the plasma sheet region. Here the magnetic field is a stretched dipole with considerable irregular "noise" in the components (*6*). The lobe consists of significantly lower ion and electron intensities than in the plasma sheet and fewer ion/electron intensity variations (although these fluctuations seem to increase as Cassini moves beyond Titan's orbit). The field has a nearly radial orientation and less noise in field components, with Bz being the smallest component (SZS coordinates) (*6*). Frequent small nondispersive changes in ion and electron intensities are consistent with particle intensity gradients (away from the center of the plasma sheet) that are produced by small-scale plasma sheet flapping, thickening, or thinning.

The LEMMS measurements of ion and electron intensities are made in situ, so that temporal and spatial variations cannot be unambiguously distinguished from one another. However, the discovery of ENA emissions from Saturn by Voyager 1 (*7*) and at Earth (*8*) enabled a new approach. The INCA sensor is designed to provide a spatially resolved map of ENA emissions over a large region, similar to the IMAGE mission at Earth (*9*). An image taken at ~2200 UT on 2 July 2004 during Cassini's outbound pass reveals a day-to-night asymmetry in the emission of ENA (Fig. 2). In this image, INCA pointed toward Saturn for only a few hours, but during that time the emitting region corotated from the dayside to the nightside as it increased in brightness. The emission region extends beyond the orbit of Titan, as is also evident in the LEMMS measurements (Fig. 1). The nightside source varies with a time scale of ~11 hours, whereas the dayside seems relatively constant. Continuous measurements over longer periods show that an ~11-hour periodicity characterizes ENA emissions from the magnetosphere (*10*).

There are ENA emissions well off the limb of Saturn, emanating from the main radiation belt and extending tailward out to the edge of the instrument's field of view (Fig. 3A). There

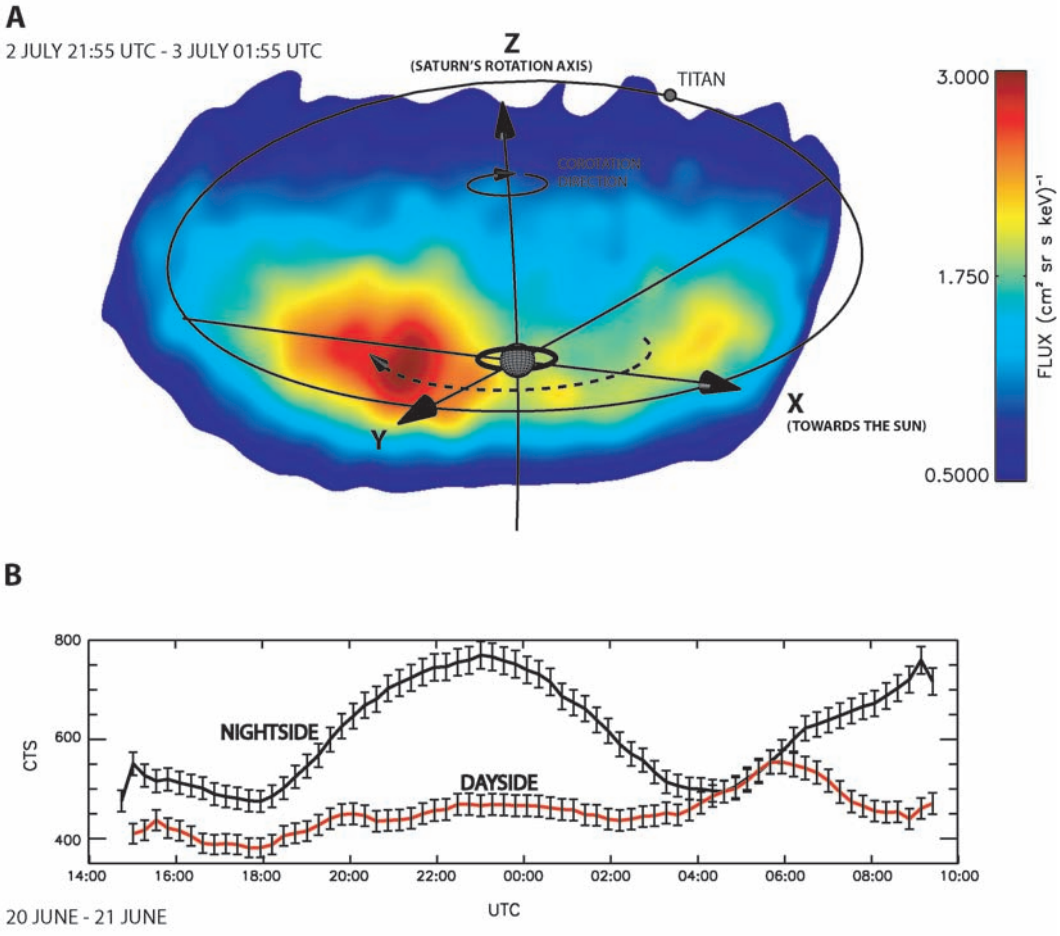


Fig. 2. (**A**) ENA image of Saturn's magnetosphere obtained by INCA on 2 July 2004, 21:55 UTC (20 to 50 keV H, 4-hour integration). Cassini was moving below the ring plane, looking back and up at Saturn. The bright emission moves around the planet with the corotational speed (dashed line). (**B**) During the approach to Saturn (20–21 June), INCA revealed 11-hour variations in the ENA emissions coming from the magnetosphere of Saturn. The rotational period of Saturn is about 11 hours. The red and black curves show temporal variation of the summed ENA counts (3 to 80 keV H, 1-hour integration) on the dayside and nightside intensities, respectively. The differences in phase and intensity in the red and black curves imply that injected plasma clouds corotate.

is also a high-intensity emission on Saturn's disk from a low-altitude equatorial belt. The ENA emission extends to a latitude 20° north of Saturn's equator. This equatorial emission originates as a belt inward of the innermost D ring, whereas the emission off the limb originates from the portion of the main radiation belt on the opposite side of Saturn (Fig. 3B).

The CHEMS instrument performed continuous measurements of the composition and charge state of ions throughout SOI, providing unique information on the likely origin of these accelerated ions in the magnetosphere. CHEMS data (Fig. 4) show that the principal constituents of the charged particle population are molecular and atomic ions, including H_2^+, O^{2+}, O^+, OH^+, H_2O^+, and O_2^+, most of which can originate as products of water dissociation. H^+ was abundant and could also originate from the dissociation of water or from Saturn's ionosphere or Titan's exosphere. He^+, which is abundant in interplanetary space (*11*), was also detected and probably enters the magnetosphere with relative ease. Most important, the relatively low concentration ($N^+/O^+ < 0.05$) of N^+—the most abundant element in Titan's atmosphere—suggests that little N_2 gas escapes from Titan's atmosphere, contrary to expectations (*12*, *13*).

The overall features of the energetic particle populations in the magnetosphere are similar to those observed during the Voyager 1 and 2 encounters more than two decades ago. There is a low-energy ion population associated with the dayside magnetopause extending inward to ~15 R_S [previously labeled "mantle" (*3*)], a rapid buildup in the intensity of ions and electrons just outside Rhea's L shell, a depletion of the ions and some electrons between the L shells of Dione and Enceladus inbound and outbound, and a rapid increase in the intensity of ions and electrons inside the orbit of Enceladus and Mimas (Fig. 1). The depletion of ions between the L shells of Dione and Enceladus suggests that the neutral gas that has been detected there for the past 6 months (*14*) may be the source for energetic particle loss through charge exchange. If so, a comparison of absolute fluxes between Cassini and Voyager data may allow us to determine the differences in neutral gas densities during the different spacecraft encounters.

Superimposed on the general morphology of the magnetosphere are some injections of plasma such as the one seen in the electrons on 2 July 2004, ~2030 UT (Fig. 1). Such injection events, characterized by energy dispersion (different energies arriving at the spacecraft at different times), have been observed in Earth's and Jupiter's magnetospheres (*15*). Injections are sudden planetward motions of hot plasma over a restricted range of longitudes and are a consequence of dynamical activity in planetary plasma environments such as in the magnetospheric substorm phenomenon. Conjectures about the existence of such injections in Saturn's magnetosphere were considered because of differences in intensities observed between the two Voyagers after successive encounters with Saturn in 1980 and 1981 (*16*).

The day-night asymmetry (Fig. 2) may be due to a larger accumulation of neutral gas on the nightside of Saturn. A maximum in particle intensity at some longitude corotating with the planet would produce more ENA on the nightside once per planetary rotation if the gas density were also maximized there. However, ENA production depends on the product of neutral gas density and particle intensity, so the excess emission could also be produced by higher nightside ion intensities. Constant particle intensity would result in bright ENA emission on the nightside but would not result in an 11-hour intensity variation of the nightside source (Fig. 2B). Hence, the interpretation of the 11-hour variation is not unambiguous at present.

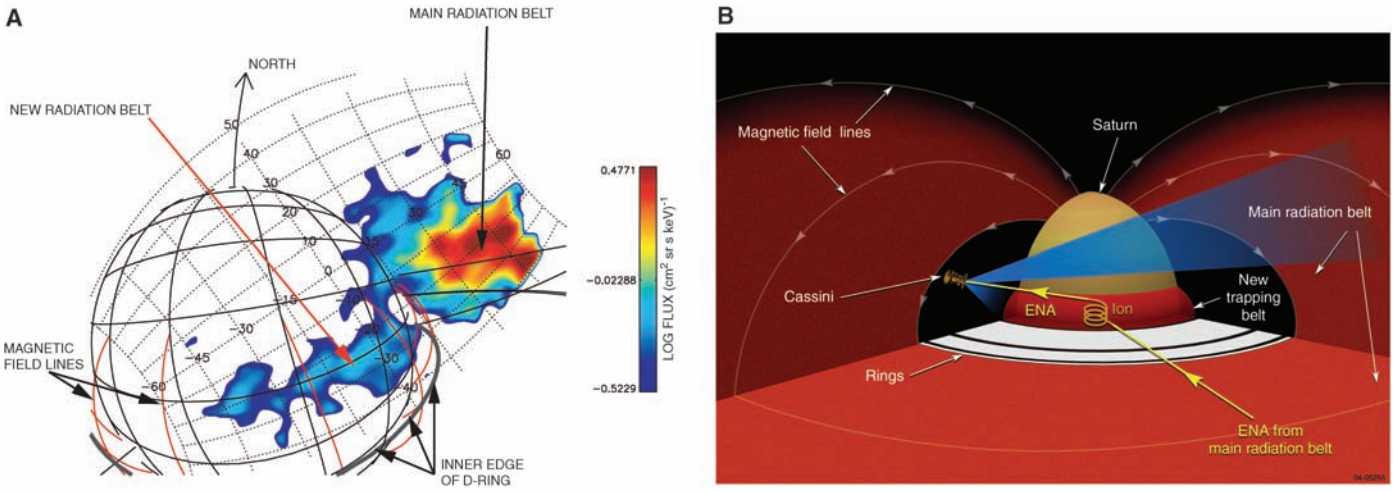


Fig. 3. ENA image obtained during Cassini's passage over the rings and interpretive schematic. (**A**) INCA image in 20 to 50 keV/nucleon ENAs. The bright region above Saturn's limb is caused by ENAs produced by charge-exchange collisions between main radiation belt ions and near-equatorial gas distributions. The band of emission above the equator on the disk is produced by the same ENAs from the main belt being stripped in Saturn's exosphere between the inner edge of the D ring and the cloud tops, trapped there temporarily as energetic ions, and then reemitted as ENAs that INCA can image. Note that the image is bounded by the limits of the INCA field of view (dotted arcs). (**B**) Schematic of the charge-exchange/stripping process that begins as ENA emission from the main belt and produces ENA emission from Saturn's exosphere.

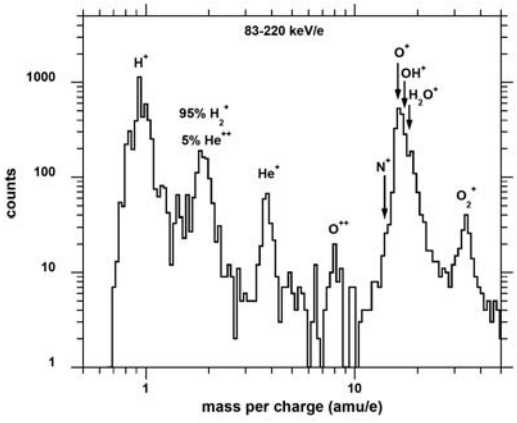


Fig. 4. Histogram of counts versus mass/charge ratios from MIMI and CHEMS, summed over the range 5.7 to 12.6 in the magnetic shell parameter L, during Cassini's outbound traversal (2 July 2004, 0800 to 1800 UTC) of the magnetosphere. The relative abundances are qualitatively correct. Although N^+ is not resolved in this display, fitting of the data shows an upper limit of $N^+/O^+ \approx 0.05$. At a mass/charge ratio of 2, the largest contribution is H_2^+ with a small amount (~5%) of He^{2+}.

The ENA-emitting region inward of the innermost D ring (Fig. 3) is explained as follows. The planet-directed ENAs from the main radiation belt are stripped of electrons when they enter Saturn's exosphere and are trapped as ions. However, the trapped ions will subsequently undergo a charge-exchange collision with exospheric atoms and be transformed back into ENAs. This process of stripping and charge exchange may be repeated many times, but some of these particles will eventually escape the exosphere as ENAs. Thus, a double charge-exchange process forms a low-altitude ENA emission region (Fig. 3B). A similar trapping belt was identified and explained by charge exchange in Earth's radiation belt (*17*) by in situ measurements at low altitudes. Although such a process at other planets might have been expected, none foresaw this happening at Saturn. Energetic particles, with their ability to pass back and forth between charged and neutral states, are temporarily unconstrained by (for instance) the Stoermer cutoff or ring absorption and therefore populate regions of space previously believed to be prohibited to them. The longitudinal extent of the inner belt has not been determined, but the ENA emission does extend all the way to the sunward edge of the INCA field of view. The ENA reemission process must be quite efficient, because the peak intensity of the low-altitude emission is about one-fifth of that from the main radiation belt.

Finally, the presence of a host of ions measured by CHEMS, apparently originating from H_2O, confirmed previous predictions (*16*, *18*) of abundant water products emanating from the rings and the icy satellites. The underlying chemistry has been described in the literature (*19*). The absence of significant amounts of N^+ ($\lesssim$5% of O^+) or N_2^+ is perplexing because N_2^+ was expected to be picked up by the magnetic field as it draped itself around Titan's exosphere (*12*, *20*). Molecular hydrogen could be a dissociation product of CH_4 in Titan's atmosphere, but the absence of nitrogen makes it more likely that the H_2^+ originated from water. H_2 at much higher energies was detected by Voyager and was thought to be escaping from Saturn's ionosphere (*3*, *4*). O_2^+ is not a direct dissociation product of water and may result from surface-gas reactions on ring particles.

References and Notes

1. J. A. Van Allen, M. F. Thomsen, B. A. Randall, R. L. Rairden, C. L. Grosskreutz, *Science* **207**, 415 (1980).
2. W. Fillius, W. H. Ip, C. E. McIlwain, *Science* **207**, 425 (1980).
3. S. M. Krimigis *et al.*, *Science* **212**, 225 (1981).
4. S. M. Krimigis *et al.*, *Science* **215**, 571 (1982).
5. S. M. Krimigis *et al.*, *Space Sci. Rev.* **114**, 233 (2004).
6. M. Dougherty *et al.*, *Science* **307**, 1266 (2005).
7. E. Kirsch, S. M. Krimigis, W. H. Ip, G. Gloeckler, *Nature* **292**, 718 (1981).
8. E. C. Roelof, D. G. Mitchell, D. J. Williams, *J. Geophys. Res.* **90**, 10991 (1985).
9. D. G. Mitchell *et al.*, *Geophys. Res. Lett.* **28**, 1151 (2001).
10. S. M. Krimigis *et al.*, 36th Division of Planetary Sciences Meeting, Louisville, KY, 8 to 12 November 2004.
11. G. Gloeckler *et al.*, in *The Heliosphere Near Solar Minimum*, A. Balogh, R. G. Marsden, E. J. Smith, Eds. (Springer-Praxis, Chichester, UK, 2001), pp. 287–326.
12. A. Eviatar, M. Podolak, *J. Geophys. Res.* **88**, 833 (1983).
13. J. D. Richardson, *Rev. Geophys.* **36**, 501 (1998).
14. L. Esposito *et al.*, *Science* **307**, 1251 (2005).
15. B. H. Mauk, D. J. Williams, R. W. McEntire, K. K. Khurana, J. G. Roederer, *J. Geophys. Res.* **104**, 22759 (1999).
16. S. M. Krimigis *et al.*, *J. Geophys. Res.* **88**, 8871 (1983).
17. J. Moritz, *Z. Geophys.* **38**, 701 (1972).
18. A. F. Cheng, S. M. Krimigis, in *Outstanding Problems in Solar System Plasma Physics: Theory and Instrumentation*, J. H. Waite, R. Moore, Eds. (American Geophysical Union, Washington, DC, 1989), pp. 253–260.
19. W. H. Ip, *Adv. Space Res.* **12**, 73 (1992).
20. H. T. Smith, R. E. Johnson, V. I. Shematovich, *Geophys. Res. Lett.* **31**, L16804 (2004).

21. The MIMI program at the Applied Physics Laboratory, Johns Hopkins University, was supported by the NASA Office of Space Science under Task Order 003 of contract NAS5-97271 between the NASA Goddard Space Flight Center and Johns Hopkins University and by subcontract at the universities of Maryland, Kansas, and Arizona.

4 October 2004; accepted 30 November 2004
10.1126/science.1105978

Composition of Saturnian Stream Particles

Sascha Kempf,[1]* Ralf Srama,[1] Frank Postberg,[1] Marcia Burton,[2] Simon F. Green,[3] Stefan Helfert,[1] Jon K. Hillier,[3] Neil McBride,[3] J. Anthony M. McDonnell,[3] Georg Moragas-Klostermeyer,[1] Mou Roy,[2] Eberhard Grün[1,4]

During Cassini's approach to Saturn, the Cosmic Dust Analyser (CDA) discovered streams of tiny (less than 20 nanometers) high-velocity (~100 kilometers per second) dust particles escaping from the saturnian system. A fraction of these impactors originated from the outskirts of Saturn's dense A ring. The CDA time-of-flight mass spectrometer recorded 584 mass spectra from the stream particles. The particles consist predominantly of oxygen, silicon, and iron, with some evidence of water ice, ammonium, and perhaps carbon. The stream particles primarily consist of silicate materials, and this implies that the particles are impurities from the icy ring material rather than the ice particles themselves.

Tiny, micrometer-sized dust particles are the building blocks of Saturn's diffuse E ring (the largest planetary ring) and are also found within Saturn's main rings. These particles provide information about the dynamics of the E ring particles and about the composition of the ring material and the surfaces of the icy moons. The Cosmic Dust Analyser (CDA) will map the composition and size distribution of the dust within the saturnian system over the course of the 4-year Cassini mission. These data will help to answer questions about the interrelation of the diffuse rings and embedded moons, ring dynamics, and the interaction between the dust and Saturn's plasma environment.

Beginning in January 2004, when the Cassini spacecraft was within 1200 R_S (Saturn radius R_S = 60,330 km), the CDA sensor discovered a few streams of tiny high-velocity dust particles (*1*) (Fig. 1). A detailed analysis of the impact signals revealed that the particles were expelled from within the saturnian system at speeds exceeding 100 km s^{-1} (*1*). The mechanism of accelerating the particles inside Saturn's magnetosphere is similar to the process that generates jovian stream particles (*2*–*4*): Positively charged grains outside the synchronous radius at 1.86 R_S will be driven outward by the electric field produced by Saturn's rotating magnetic field.

Because of its field, Saturn's magnetosphere acts like a mass spectrometer, allowing only particles within a certain mass and velocity range to escape. The dynamical properties of the earliest detected particles (speed > 100 km s^{-1} and radius 10 to 20 nm) suggested that the outskirts of Saturn's dense A ring were the most probable dust source.

The CDA (*5*) consists of two independent dust detection systems: the Dust Analyser (DA), which is sensitive to particles within a large mass (5×10^{-18} to 1×10^{-12} kg) and velocity (1 to 100 km s^{-1}) range, and a High Rate Detector (HRD), which is a polyvinylidene fluoride foil sensor for monitoring high impact rates up to 10^4 s^{-1} in dust-rich environments. The DA subsystem simultaneously measures the charge carried by the dust grain, its mass, its impact speed, and its elemental composition. This is accomplished by a combination of three sensors: an impact ionization detector similar to the instruments on the Galileo and Ulysses spacecraft (*6*); a charge-sensing unit; and a time-of-flight (TOF) mass spectrometer, the Chemical Analyzer (CA).

The CA operates by separating the plasma constituents (from the particle and the Rh target on the instrument) into ions and electrons in a strong electric field. Ions are accelerated toward an electron multiplier, producing a TOF mass spectrum. In the ideal case, the ion's flight time t^i is a

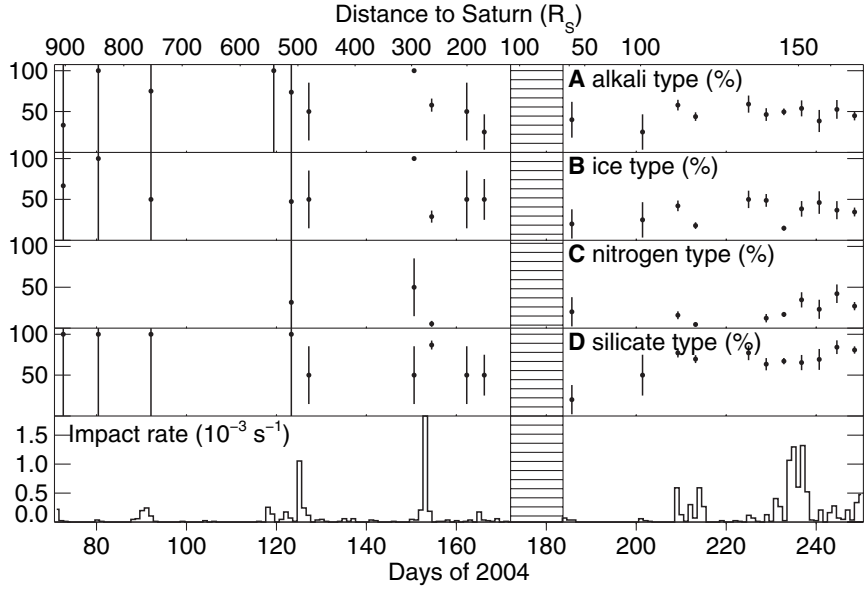


Fig. 1. Stream particle impact rates registered by the CDA during Cassini's approach to Saturn and during the outbound leg of Cassini's first orbit around Saturn (lower panel). Between 20 June 2004 (day 172) and 1 July 2004 (day 183), the CDA detector was powered off (marked by the outlined area). Cassini entered Saturn's magnetosphere on 28 June (day 180) and left the magnetosphere on 5 July 2004 (day 187). The CDA was not continuously geometrically sensitive to stream particles. After Cassini's SOI, most of them were detected during the daily 9-hour data transmission to Earth, when Cassini slowly rotated about its antenna axis. The sequence of data transmissions was almost evenly distributed, so that the monitored impacts reproduce the stream particle flux on time scales longer than 12 hours. For Cassini's approach phase, this was unfortunately not the case (*1*). The upper panels display the percentage of spectra showing lines typical of alkali elements (**A**), ice (**B**), nitrogen compounds (**C**), and silicates (**D**) as a function of time. The vertical bars indicate the uncertainty associated with counting statistics.

[1]Max-Planck-Institut für Kernphysik, Saupfercheckweg 1, 69117 Heidelberg, Germany. [2]Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, CA 91109, USA. [3]Planetary and Space Sciences Research Institute, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK. [4]Hawaii Institute of Geophysics and Planetology, University of Hawaii, 1680 East West Road, Honolulu, HI 96822, USA.

*To whom correspondence should be addressed. E-mail: Sascha.Kempf@mpi-hd.mpg.de

direct measure of its mass m, although for this type of instrument, the resulting line shapes are affected by the ions' initial energy distribution.

From the first detection of saturnian stream particles in January 2004 until Cassini's arrival at Saturn at the end of June 2004, a total of 76 mass spectra were recorded. In addition, during the outbound leg of Cassini's first orbit about Saturn, another 508 mass spectra were registered.

The mass and speed of the saturnian stream particles based on theoretical calculations are outside the calibrated range of the detector (*1*). Thus, these parameters cannot be derived from the impact signals. The amount of plasma generated by the impact, however, should scale with the particle mass and impact speed. Particles detected outside about 500 R_S caused relatively large impact charges, which supports an A-ring origin, whereas particles registered after Cassini's Saturn orbit insertion (SOI) generally produced little impact plasma, which is dynamically compatible with an E-ring origin (*1*). With the exception of four particles detected soon after the SOI, all stream particles were registered outside Saturn's magnetosphere, although their source lies within. All mass spectra show distinct mass lines of mostly atomic ions in the lower mass range [n < 150 atomic mass units (amu)], which is indicative of high impact energies. The spectra are dominated by rock-forming elements such as O and Si, by the target element Rh, and by C of uncertain origin. This is in contrast to spectra of jovian stream particles registered by the CDA in 2000, which are dominated by C^+, Na^+, and K^+ but show little O^+ (bottom panel of Fig. 2).

Most of the spectra are composed of less than about 30,000 ions. For those impacts, the recording of the event is not triggered by the target signals but by the H^+ line exceeding the multiplier threshold. In such a case, the offset time t^0 for assigning a mass scale is larger than for events triggered by one of the target signals. The H^+ comes from the Rh target. Because it takes little energy to release the H^+ into the impact plasma cloud, even impacts by tiny particles will generate an H^+ line strong enough to trigger the event recording.

A reliable identification of mass lines within the faint noisy spectra is a challenging task. In some spectra, only a single well-resolved line is apparent (Fig. 2). For that reason, we used an iterative method to magnify features too weak to be recognized easily in some of the individual spectra. We identified the most prevalent mass lines—C^+, O^+, Na^+, K^+, Fe^+, SiO_2^+, and Rh^+—by adding spectra of similar strength and with similar features. Then we scanned each individual spectrum for these lines to improve the mass scale calibration. This procedure was repeated until the resolution of the spectral lines of the sum spectrum did not improve. Finally, we obtained a consistent data set of calibrated individual mass spectra. Besides the aforementioned ions, we found evidence for H_3O^+, which is characteristic of water ice. There is no indication, however, of water cluster ions $(H_2O)_xH_3O^+$, with x = 1, 2, . . ., typically found for impacts by water ice grains at low impact energies (*7*). About 25% of the spectra also showed N^+ and a line at 18 amu, which may be NH_4^+. Likely candidates for the frequent strong line at 28 amu are CO^+ and Si^+. There is evidence of SiO_2^+ and Fe^+ in our data, which are building blocks of typical silicate minerals such as olivine $[(Mg,Fe)_2SiO_4]$. Laser ionization experiments with olivine, serpentine, and Murchison meteorite performed to simulate impacts at speeds between 23 and 65 km s^{-1} yielded Fe^+ and SiO_2^+ lines (*8*). There was also an indication of a weak 28-amu line in the laser experiments, which was attributed to Si^+. At stream particle impact speeds ≥100 km s^{-1}, the number of Si^+ ions will be larger than in the laser experiments. In addition, the 60-amu line attributed to SiO_2^+ correlates with the 28-amu line. Thus, assigning the 28-amu line in our data to Si^+ seems to be the more natural assumption.

Low-energy impacts are known to often form target-projectile clusters built up of the target atom and the most abundant low-mass projectile atom or molecule (*9*, *10*). Even target-projectile compounds that are difficult to form, such as $RhAl^+$ or $RhFe^+$, showed up in mass spectra taken in impact experiments (*10*). Thus, target-projectile ions are indicators of prominent low-mass projectile substances, provided that the corresponding projectile ion shows up in the spectrum as well. At higher impact energies, target-projectile ions are less likely to form and will produce only faint lines, if any. We found a line at 28 amu above the Rh peak (Fig. 3), which we identify as a target/projectile cluster formed of Rh and Si. From this we conclude that Si^+ is a major stream particle component.

All the spectra are dominated by a C^+ line, although it is unclear how much C originates in the particle. If all of the C is due to the particle, it must constitute the majority of the mass. This feature resembles the CHON-component found in impact spectra of dust particles from comet Halley (*11*). However, C appears in all calibration spectra, even when there is no C present in the impactor.

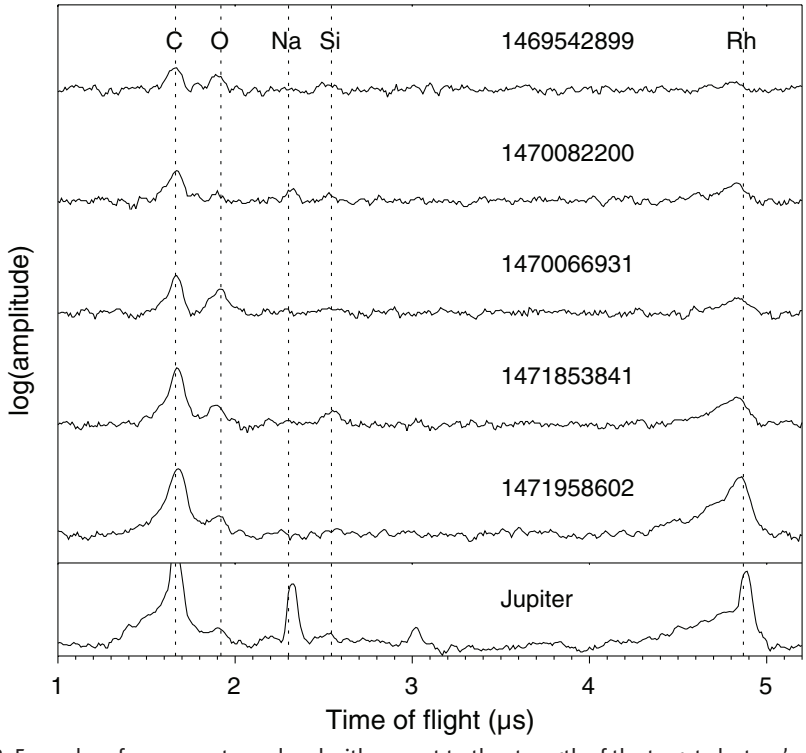


Fig. 2. Examples of mass spectra ordered with respect to the strength of the target electron's signal. The uppermost spectrum (1469542899) is composed of less than 18,000 ions, whereas about 180,000 ions contributed to the lowest saturnian spectrum shown (1471958602). The dotted lines indicate the masses of C, O, Na, Si, and the target material Rh. The dominant ions are C and Rh. The shift of the Rh peak mass with an increasing ion signal is a phenomenon also known from laboratory experiments. For comparision, the bottom panel shows a typical spectrum of a jovian stream particle, which is dominated by C^+, Na^+, K^+, and Rh^+.

To characterize the spectra qualitatively, we assigned to each individual spectrum, according to its identified lines, up to four "composition types": a silicate type to spectra showing at least an Si^+, SiO_2^+, or Fe^+ line; an alkali type to spectra showing at least an Na^+ or K^+ line; a water ice type to spectra showing an O^+ and H_3O^+ line; and a nitrogen type to spectra showing an N^+ and NH_4^+ line. About 12% of the spectra could not be associated with one of the composition types, and 74% of the spectra were of the silicate type (32% of them simultaneously belonged to the ice type), but only about 7% of the spectra were of the ice type but showed no silicate lines. The preponderance of spectra with silicate lines suggests that most of the detected grains consist of a silicate material. If the water ice lines were not due to target contamination, then about a third of the silicate grains had ice mantles, whereas solid ice grains were rare.

The likelihood of finding a line in an individual spectrum depends on the signal-to-noise ratio, which itself is related to the amount of plasma q_e generated by the impact. Fluctuations in the occurance of the less abundant composition types (ice and nitrogen) correlate with periods during which only faint impacts were registered. However, the dominant silicate and alkali composition types show no pronounced time dependence (no variation with saturnian distance). This suggests that the time dependence of the ice and nitrogen types is at least partially due to the data processing. Taking this into account, we conclude that the source of the saturnian stream particles needs to be fairly homogeneous. Particles originating from different locations within the A ring and E ring are made of similar material. About 50% of the spectra show at least one alkali line. Moreover, there is a correlation between the detection probabilities for Na^+ and K^+. A TOF–secondary mass spectroscopy analysis of the CDA Rh target sample shows that there are stochastically distributed small patches of Na and K impurities beyond the surface, which total about 54% of the target area (*12*). This implies that the alkali elements stem from the target.

Saturn's main ring as well as its E ring are thought to be primarily made of water ice, possibly containing clathrate hydrates of ammonia or methane and a minor amount of impurities, most likely iron-bearing silicate compounds (*13*). This idea is consistent with our findings.

Although N_2 (42%), H_2O (29%), CO (17%), and CO_2 (9%) are the major combustion products of the bipropellant fuel monomethylhydrazene and nitrogen tetroxide burnt by Cassini's main engine (*14*), there is no evidence of contamination from this source. First, there was no significant increase in the share of icy and nitrogen spectra after the 1.5-hour main engine burn during SOI on 30 June 2004 (Fig. 1). Second, 8-hour heat-ups of the CA target on 21 March 2004 and 10 October 2004 to remove possible contamination did not alter the frequency of those spectral lines.

Our findings indicate that most of the stream particles detected by the CDA are composed of a silicate mineral, at least 30% of which give spectral signatures of water ice and/or clathrate hydrates of ammonia (which may be due to icy mantles). We suggest that stream particles were formed either by collisions between the parent ring bodies (for grains originating from the main rings) or by collisions of ring particles with an icy moon (in the case of an E-ring origin). In both cases, small silicate impurities embedded within the ice are likely to survive the collision, eventually forming the stream particle reservoir. Our data do not provide evidence of solid ice grains smaller than about 20 nm. Furthermore, the mass spectra imply that the composition at least of the particles smaller than ~20 nm is essentially the same. Thus, the reservoir of stream particles capable of escaping from the saturnian system essentially consists of the silicate impurities of the ring material. In situ studies of stream particles allow investigation of this material, which is not accessible by other means. A likely source of these impurities is probably the continuous stream of dust particles originating in the Kuiper belt, as well as interstellar particles traversing the solar system.

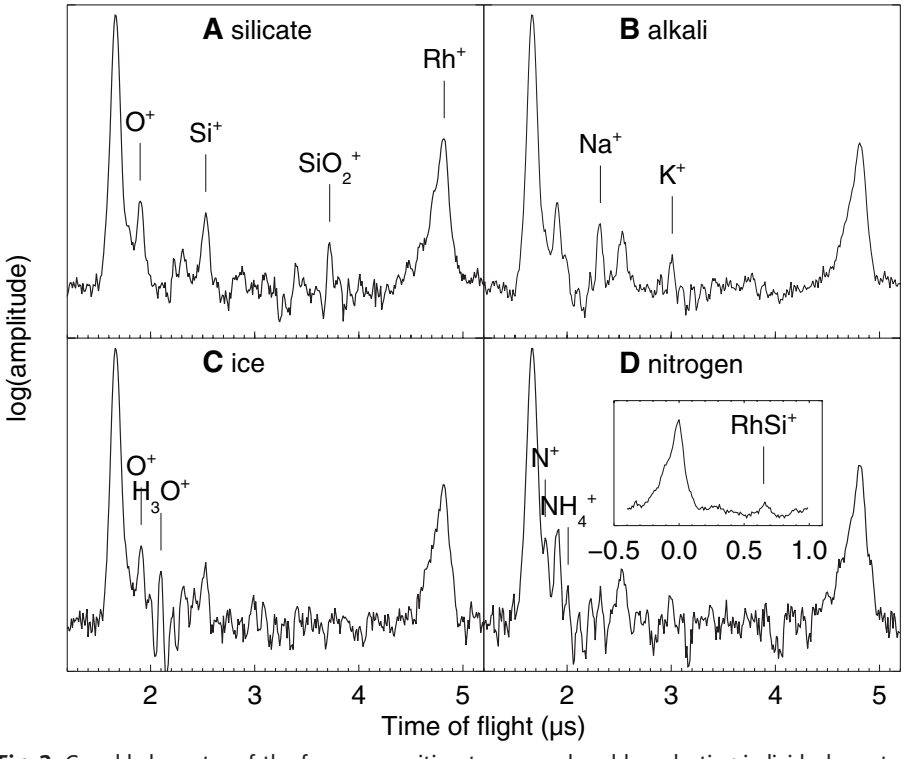


Fig. 3. Co-added spectra of the four composition types, produced by selecting individual spectra with $q_e < 15$ femto-Coulomb (fC) and showing Si^+, SiO_2^+, or Fe^+ [(**A**), silicate]; Na^+ or K^+ [(**B**), alkali]; O^+ and H_3O^+ [(**C**), ice]; and N^+ and NH_4^+ [(**D**), nitrogen]. In each case, features of the other types are still apparent. The inset in (D) shows the RhSi cluster-ion line obtained by summing up all spectra with a clear Rh^+ line and $q_e < 15$ fC relative to the Rh peak position (in microseconds).

References and Notes

1. S. Kempf *et al.*, *Nature* **433**, 289 (2005).
2. M. Horányi, G. Morfill, E. Grün, *J. Geophys. Res.* **98**, 21245 (1993).
3. D. Hamilton, J. Burns, *Nature* **364**, 695 (1993).
4. M. Horányi, *Phys. Plasmas* **7**, 3847 (2000).
5. R. Srama *et al.*, *Space Sci. Rev.* **114**, 465 (2004).
6. S. Auer, in *Interplanetary Dust* (Springer, Berlin, 2001), pp. 387–438.
7. T. Timmermann, E. Grün, in *Origin and Evolution of Dust* (Kluwer, Dordrecht, Netherlands, 1991), pp. 387–438.
8. T. Ahrens, S. Gupta, G. Jyoti, J. Beauchamp, *J. Geophys. Res.* **108**, 1-1 (2003).
9. W. Knabe, thesis, University of Heidelberg, Heidelberg, Germany (1983).
10. M. Stübig, thesis, University of Heidelberg, Heidelberg, Germany (2002).
11. E. Jessberger, J. Kissel, *Chemical Properties of Cometary Dust and a Note on Carbon Isotopes* (Kluwer, Dordrecht, Netherlands, 1991), pp. 1075–1092.
12. A. Mocker, F. Postberg, S. Kempf, unpublished observation.
13. L. Esposito *et al.*, *Saturn* (Univ. of Arizona Press, Tuscon, AZ, 1984), pp. 463–545.
14. C. Guernsey, personal communication.
15. We thank E. Jessberger and D. Rost for analyzing the composition of the CA target. The group at the Max-Planck-Institut für Kernphysik was supported by the Max-Planck Society and by the Deutsches Zentrum für Luft- und Raumfahrt under grant 500OH9802.

8 October 2004; accepted 10 January 2005
10.1126/science.1106218

Bacterial Injectisomes: Needle Length Does Matter

Luís Jaime Mota, Laure Journet, Isabel Sorg, Céline Agrain, Guy R. Cornelis*

Type III secretion (T3S) allows Gram-negative pathogenic bacteria adhering to the membrane of a eukaryotic cell to paralyze or reprogram this cell by injecting proteins into its cytosol (*1*). Many T3S nanomachines (injectisomes) possess a stiff needlelike structure of a defined length (*2*). The needle is thought to function as the conduit for protein translocation. In *Yersinia* bacteria, including *Y. pestis*, the needle length is defined by the protein ruler YscP (*3*), and the proteins injected upon host cell contact, called Yops, are involved in, among other things, caspase activation and macrophage apoptosis (*4*).

To address the question of why the needle length is controlled, we replaced *yscP* (515 codons) on the 70-kb virulence plasmid of *Y. enterocolitica* E40 by either a truncated (388 codons) or an enlarged (680 codons) allele. We incubated the modified bacteria in conditions that artificially induce *Yersinia* T3S (*4*) and observed no clear difference in Yop secretion in relation to wild-type bacteria (fig. S1). We compared *Yersinia* that make short ($YscP_{388}$, needle length L = 45 ± 13 nm) and long ($YscP_{680}$, L = 88 ± 23 nm) needles to wild-type bacteria (L = 55 ± 11 nm) (fig. S2) for their capacity to inject the apoptosis-inducing YopP into J774A.1 macrophages, by assaying caspase activation in infected cells. Bacteria making long needles were as efficient as the wild type, but those making short needles were not (Fig. 1A). This suggested that the needle must span a minimal distance, presumably determined by the *Y. enterocolitica* adhesins YadA and invasin (Inv).

Although YadA protrudes ~28 nm from the bacterial outer membrane (*5*), Inv is only 18 nm long (*6*). YadA (455 amino acids) has a lollipop-shaped structure with an oval head on top of a coiled-coil rod (*5*). Variations in coiled-coil domain length result in longer or shorter YadAs (*5*), modulating the distance between the bacterium and host-cell lipid membranes. We altered YadA length by deletion or duplication of residues 230 to 319 within the coiled-coil rod, yielding shorter ($YadA_{365}$, estimated length of 15 nm) and longer ($YadA_{545}$, ~40 nm) adhesins. We disrupted *inv* and *yadA* in *Yersinia* carrying $yscP_{388}$, $yscP_{wt}$, and $yscP_{680}$ and introduced plasmids encoding $YadA_{wt}$, $YadA_{365}$, and $YadA_{545}$. The YadA variants were as functional as $YadA_{wt}$ in promoting cell attachment (fig. S3). Caspase activity in *Yersinia*-infected macrophages revealed that shortening YadA suppressed the defect of short needles, whereas lengthening YadA enhanced the deficiency (Fig. 1A). Accordingly, lengthening YadA reduced the efficacy of wild-type needles but not of long needles (Fig. 1A). We monitored the translocation of YopO into J774A.1 by using an adenylate cyclase (Cya) reporter assay (Fig. 1B), confirming that short needles lead to a defect in Yop injection.

Thus, increasing the distance between the needle tip and the host cell by shortening the needle or by lengthening YadA reduces translocation; i.e., the needle needs to have a minimal length to be fully functional. We tested if bacteria with impaired translocation efficiency secreted Yops into the medium when incubated with J774A.1. Regardless of the length of the needle or YadA, we could not detect YopE in this fraction (Fig. 1C), indicating that there was no leakage and suggesting that contact between the needle tip and the host cell membrane is necessary to trigger T3S. It is thus likely that needle length has evolved to match specific structures at the bacterial and host cell surfaces.

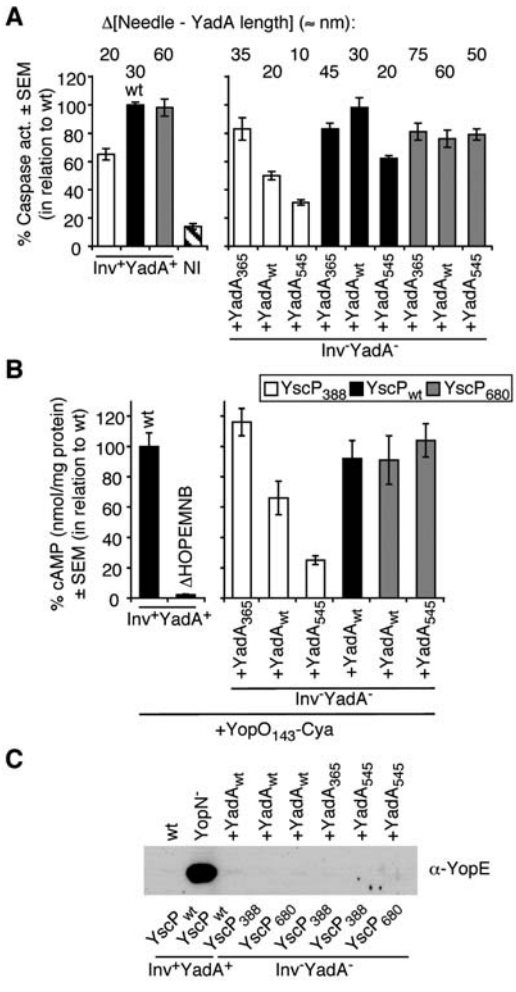


Fig. 1. (**A**) Caspase activity (act.) on *Y. enterocolitica* E40-infected macrophages. Results are the mean ± SEM from five independent experiments, each done in triplicate. The numbers on top indicate how far the needle sticks out (in nm) relative to YadA. NI, not infected; wt, wild-type. (**B**) $YopO_{143}$-Cya translocation into macrophages. ΔHOPEMNB is a negative control, not expressing the YopB translocator (*4*). Data are the mean ± SEM from three independent experiments, each done in triplicate. cAMP, cyclic adenosine monophosphate. (**C**) Immunoblot of cultured supernatant proteins from *Y. enterocolitica* E40–infected macrophages. YopN– is a positive control; in a *yopN* mutant, Yop translocation is independent of contact (*4*).

References and Notes

1. J. E. Galán, A. Collmer, *Science* **284**, 1322 (1999).
2. T. C. Marlovits *et al.*, *Science* **306**, 1040 (2004).
3. L. Journet, C. Agrain, P. Broz, G. R. Cornelis, *Science* **302**, 1757 (2003).
4. G. R. Cornelis, *Nature Rev. Mol. Cell Biol.* **3**, 742 (2002).
5. E. Hoiczyk, A. Roggenkamp, M. Reichenbecher, A. Lupas, J. Heesemann, *EMBO J.* **19**, 5989 (2000).
6. Z. A. Hamburger, M. S. Brown, R. R. Isberg, P. J. Bjorkman, *Science* **286**, 291 (1999).

7. We thank M. Dürrenberger, G. Morson, and U. M. Spornitz for use of electron microscope facilities; J. A. Bengoechea and U. Jenal for discussions; K. T. Hughes and J. Pieters for critical reading; and the Swiss National Science Foundation for grant no. 32-65393.01.

Supporting Online Material
www.sciencemag.org/cgi/content/full/307/5713/1278/DC1
Materials and Methods
Figs. S1 to S3
References and Notes

18 November 2004; accepted 16 December 2004
10.1126/science.1107679

Biozentrum, Universität Basel, 4056 Basel, Switzerland.

*To whom correspondence should be addressed. E-mail: guy.cornelis@unibas.ch

The Selective Cause of an Ancient Adaptation

Guoping Zhu,[1]* G. Brian Golding,[3] Antony M. Dean[1,2]†

Phylogenetic analysis reveals that the use of nicotinamide adenine dinucleotide phosphate (NADP) by prokaryotic isocitrate dehydrogenase (IDH) arose around the time eukaryotic mitochondria first appeared, about 3.5 billion years ago. We replaced the wild-type gene that encodes the NADP-dependent IDH of *Escherichia coli* with an engineered gene that possesses the ancestral NAD-dependent phenotype. The engineered enzyme is disfavored during competition for acetate. The selection intensifies in genetic backgrounds where other sources of reduced NADP have been removed. A survey of sequenced prokaryotic genomes reveals that those genomes that encode isocitrate lyase, which is essential for growth on acetate, always have an NADP-dependent IDH. Those with only an NAD-dependent IDH never have isocitrate lyase. Hence, the NADP dependence of prokaryotic IDH is an ancient adaptation to anabolic demand for reduced NADP during growth on acetate.

A central goal in evolutionary biology is the understanding of the mechanisms underlying ancient adaptive events (*1*). Guided by phylogenetic analysis, genetic engineering has been used to reconstruct ancestral genes, allowing ancient phenotypes to be revealed (*2*–*5*). Reconstruction can tell us what the ancestral phenotype was like, but it does not tell us why the modern phenotype evolved. Were the changes adaptive or just due to chance? If they were adaptive, what was the basis of that adaptation? Rigorous tests of hypotheses regarding adaptation are needed to identify the mutations responsible for changes in function and to demonstrate that they are selectively advantageous under the conditions specified (*6*). This has yet to be achieved for any reconstructed gene. The isocitrate dehydrogenase (IDH) system offers an opportunity to elucidate the adaptive basis of an ancient functional change.

Phylogenetics. IDHs belong to a large, ubiquitous, and very ancient family of enzymes that play central roles in energy metabolism (*7*, *8*), amino acid biosynthesis (*9*, *10*), and vitamin production (*11*). Although most family members use nicotinamide adenine dinucleotide (NAD) to oxidize their substrates, some IDHs, including that of *E. coli*, use NADP instead (*7*).

Phylogenetic analyses reveal that NAD use is an ancestral trait and that NADP use by prokaryotic IDH arose on or about the time that eukaryotic mitochondria first appeared, some 3.5 billion years ago (Fig. 1). Phylogenetic trees were constructed using maximum parsimony and neighbor joining, with deep branches swapped and assessed by maximum likelihood (*12*). Different phylogenetic methods often give slightly different trees, and others produce sets of slightly different trees whose likelihoods are indistinguishable. This is especially true when considering deep phylogenies whose sequences diverged from a common ancestor billions of years ago. For the ancient family of IDHs and isopropylmalate dehydrogenases (IMDHs), short branches among the deeper nodes lead to a large number of trees with similar likelihoods. Nevertheless, all trees contain three well-supported monophyletic groups: type I IDHs, the highly divergent type II IDHs, and the NAD-IMDHs and related enzymes. There is no suitable outgroup. Instead, we rooted the phylogeny on the branch linking the IDHs to the IMDHs because the common ancestor was undoubtedly prototrophic, capable of synthesizing glutamate and leucine, and hence must have had both enzymatic functions (*13*). This root implies that specificity toward NADP

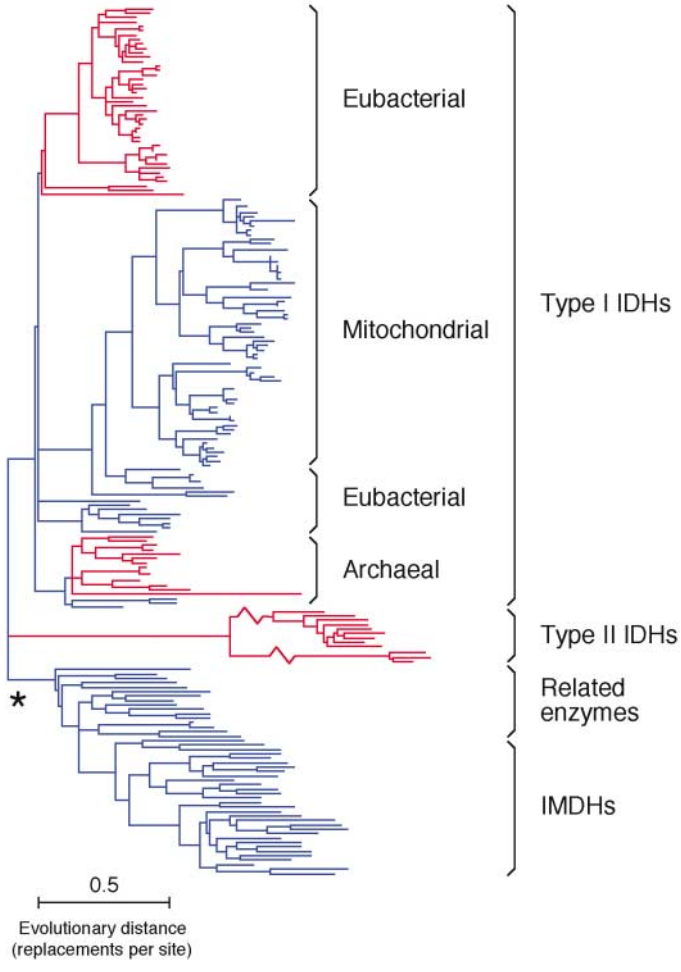


Fig. 1. A maximum likelihood phylogeny of the IDH family. This tree, like all others with similar likelihoods, reveals that NADP use evolved independently several times (red lineages) billions of years ago, around the time eukaryotic mitochondria first appeared. The asterisk represents the approximate position of the root based on biochemical evidence (*13*).

[1]BioTechnology Institute, [2]Department of Ecology, Evolution, and Behavior, University of Minnesota, St. Paul, MN 55108, USA. [3]Department of Biology, McMaster University, Hamilton, Ontario L8S 4K1, Canada.

*Present address: Department of Biology, Anhui Normal University, Wuhu 241000, People's Republic of China.

†To whom correspondence should be addressed. E-mail: adean@biosci.umn.edu

arose in prokaryotes several times, on or about the time the eukaryotic mitochondria first appeared, some 3.5 billion years ago.

Recent structural comparisons (*14*) confirm that NADP use by type I IDHs evolved independently from the use of NADP by type II IDHs and that the monomeric type II IDHs (lower cluster, eubacterial) are derived from the dimeric enzymes (upper cluster, eubacterial and eukaryotic). Specificity toward NADP may have arisen twice within the type I IDHs, because the same suite of amino acid replacements for binding NADP are found in both the eubacterial and archaeal lineages. Other trees with similar likelihoods lead to similar conclusions. We conclude that the use of NADP is an ancient adaptation that arose on at least three occasions in early prokaryotes.

Structural biology. To identify the amino acids that are responsible for the change in specificity from NAD to NADP, it would be ideal to compare the structures of NADP-using with NAD-using IDHs and contrast the modes of coenzyme binding. Unfortunately, no such structures of NAD-using IDHs are available. Although divergent in sequence, IMDHs are homologous to the IDHs (Fig. 1) and structurally very similar to the IDHs, and they use NAD (Fig. 2A). Amino acid residues determining coenzyme use were identified from high-resolution crystallographic structures of the binary complexes of NADP bound to *E. coli* IDH (*15*) and of NAD bound to *Thermus thermophilus* IMDH (*16*) (Fig. 2A). Maximum likelihood reconstruction of ancestral sequences confirmed that all six conserved residues binding NADP in prokaryotic IDHs were introduced in very ancient times (*13*). This analysis also revealed that the three key residues conferring specificity toward NAD are ancestral and that the others are not conserved in NAD-specific family members, because they have no role in binding NAD.

Protein engineering. Protein engineering (Fig. 2B) was used to switch the coenzyme specificity of *E. coli* IDHs from NADP to NAD (*17*). That experiment showed that amino acids in the active site, and no others, determine specificity. Our engineered IDH has five replacements in the active site. Three [Lys344Asp (where Lys^{344} is replaced by Asp), Tyr345Ile, and Val351Ala] are conserved, ancestral, and confer specificity toward NAD, and two (Tyr391Lys and Arg395Ser) are not conserved and simply remove interactions with the 2′-phosphate of NADP. The engineered IDH has two additional replacements (Cys201Met and Cys332Tyr) outside the active site that increase overall activity but do not affect specificity. X-ray crystallography (Fig. 2B) shows that the engineered IDH binds NAD in precisely the same manner as wild-type IMDH (*18*). Kinetic studies (*17*) show that specificity was inverted by a factor of 1.4 million, from a 7000-fold preference for NADP ($k_{cat}^{NADP}/K_m^{NADP}$ = 4.7 × 10^6 M^{-1} s^{-1}, k_{cat}^{NAD}/K_m^{NAD} = 690 M^{-1} s^{-1}, where k_{cat} is the catalytic rate constant and K_m is the Michaelis constant) to a 200-fold preference for NAD ($k_{cat}^{NADP}/K_m^{NADP}$ = 810 M^{-1} s^{-1}, k_{cat}^{NAD}/K_m^{NAD} = 1.64 × 10^5 M^{-1} s^{-1}). With a k_{cat} = 16.2 s^{-1} and a K_m = 99 μM, our engineered IDH is as active as naturally occurring NAD-dependent members of the family (*17*).

Metabolism. Previously (*13*), we hypothesized that the use of NADP by prokaryotic IDHs is an adaptation to growth on acetate. When bacteria grow on highly reduced (energy-rich) compounds such as glucose, NADPH (the reduced form of NADP) for biosynthesis is obtained by diverting energy-rich carbon from glycolysis into the oxidative branch of the pentose phosphate pathway (*19*). During growth on acetate, which is a highly oxidized (energy-poor) compound, there is no energy-rich carbon to divert into the oxidative branch, and therefore alternative sources of NADPH are required. These are a transhydrogenase (PntAB) encoded by the gene *pntAB* (*20*), an NADP-dependent malic enzyme (MAEB) encoded by *maeB* (*21*), and the NADP-dependent IDH encoded by *icd* (*22*). Metabolic flux analyses of *E. coli* consuming acetate show that IDH provides about 90% of the NADPH (*21*, *23*–*25*).

Competition studies. Testing whether the use of NADP by eubacterial type I IDH is an adaptation to growth on acetate requires competing strains that differ only at their chromosomally located *icd* gene. Therefore, the wild-type chromosomal copy of *icd*^{NADP} was replaced with the engineered *icd*^{NAD} (*26*). Chemostat competition experiments (*27*–*29*) were conducted in a minimal salts medium with either acetate or glucose as the sole limiting nutrient. The engineered *icd*^{NAD} was strongly selected against during competition for acetate, yet favored slightly over wild-type *icd*^{NADP} during competition for glucose (Fig. 3). These results support our hypothesis that the switch in coenzyme specificity by IDH was driven by the need to supply additional NADPH for biosynthesis during growth on acetate.

Cause of selection. The selection on acetate might be attributed to the engineered IDH having altered kinetic characteristics toward isocitrate and/or being less efficiently regulated by IDH kinase/phosphatase (*30*), rather than attributed to the switch in coenzyme use. We reasoned that the selection at *icd*, if caused by the change in coenzyme use, should intensify as other sources of NADPH are removed (MAEB and PntAB), because IDH would now contribute proportionally more reducing power to biosynthesis. By contrast, selection at *icd* is not expected to intensify if the change in coenzyme use is of no functional consequence.

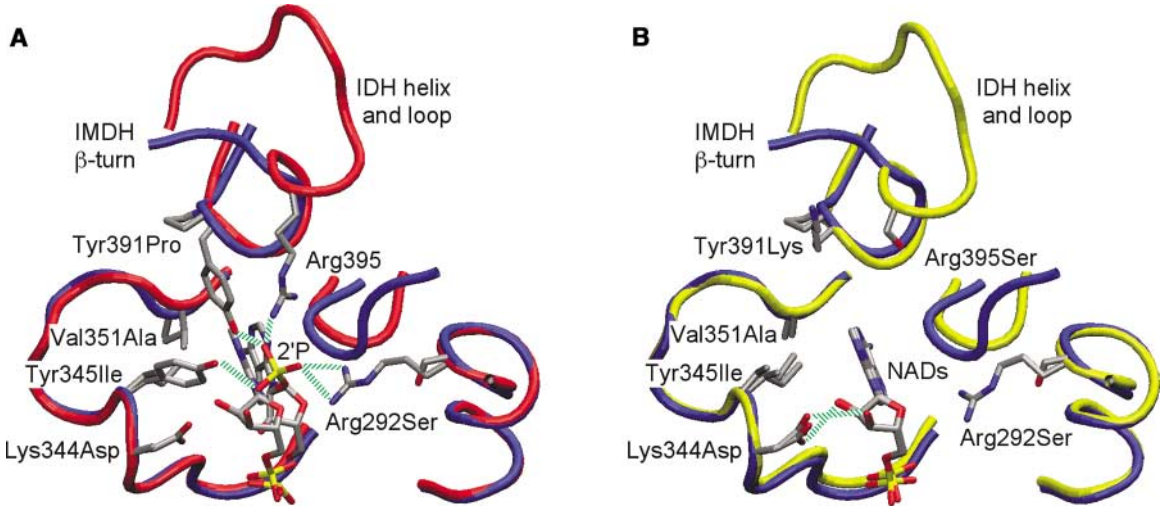


Fig. 2. Structural analysis reveals the modes of binding of NADP and NAD in the active sites of IDH and IMDH. Only key residues are shown, with side chains colored using the CPK (Corey, Pauling, and Kultun) convention (gray, carbon; red, oxygen; blue, nitrogen; yellow, phosphorus). Labels designate the IDH amino acid, the site number in *E. coli* IDH, and the IMDH amino acid. (**A**) Wild-type *E. coli* IDH (main chain red) with H bonds (green dashes) to the 2′-phosphate of NADP (2′P) superimposed on wild-type *T. thermophilus* IMDH (main chain blue) with NAD. (**B**) Engineered IDH (main chain yellow) reveals that NAD binds in precisely the same manner as seen in *T. thermophilus* IMDH and that the same H bonds form to the ribose hydroxyls.

Strains with deletions of *maeB*, *pntAB*, and *udhA* (which encodes a soluble transhydrogenase, UdhA) that are generated by the method of Datsenko and Wanner (*31*) grow so slowly on acetate relative to the wild type (most simply wash out from the chemostat) that fitnesses were best estimated from growth rates in batch cultures (*26*). Table 1 presents the growth rates of the engineered icd^{NAD} relative to icd^{NADP} in various genetic backgrounds. On acetate, selection against the engineered icd^{NAD} in Δ*pntAB* and Δ*maeB* backgrounds intensifies as the number of NADPH-producing steps declines. Selection against the engineered icd^{NAD} is much weaker in the control experiments on glucose because other pathways provide the bulk of NADPH for growth.

In Δ*udhA* backgrounds, the engineered icd^{NAD} can be favored over the wild-type allele (Table 1). UdhA is a soluble transhydrogenase that, unlike PntAB, lacks a source of energy to drive hydride transfer from one coenzyme to another (*25*). Consequently, the overall direction of the UdhA-catalyzed reaction depends on the ratios of NADH to NAD and of NADPH to NADP. In wild-type cells growing on acetate, UdhA usually favors production of NADH through the oxidation (or reoxidation) of abundant NADPH (*25*). Replacing icd^{NADP} with the engineered icd^{NAD} in a Δ*udhA* genetic background increases NADH production and improves fitness (Table 1). Deleting *pntAB* and/or *maeB* does nothing to restore NADH levels, and fitness does not improve (*32*). These results confirm that UdhA is an important source of NADH during growth on acetate, but not on glucose (Table 1). These experiments demonstrate that the engineered icd^{NAD} can provide NADH for energy in certain mutant genetic backgrounds.

These results are entirely in accord with the hypothesis that selection on acetate is caused by the change in coenzyme use and not by changes in kinetics toward isocitrate and/or changes in regulatory phosphorylation. We speculate that evolutionary functional reversals at *icd* might be possible if NADH production becomes limited (if adaptive changes in other steps reduce NADH production as a correlated response) or if demand for NADH increases because of some environmental change (for example, growth at an extreme ionic strength or extraction of scarce resources from the environment, either of which can make increased energetic demands on membrane transporters).

Genomic comparisons. Complete genome sequences provide additional support for our hypothesis. Isocitrate lyase (ICL) is essential for growth on acetate and provides carbon for biosynthesis by diverting isocitrate into the glyoxylate bypass and away from the two CO_2-releasing steps in the Krebs cycle (*7*, *19*, *21*, *23–25*). Without exception, each of the 46 prokaryotic genomes encoding an ICL has an NADP-dependent IDH, and no ICL is found in the 12 genomes encoding an NAD-dependent IDH (table S1). Although shared phylogenetic histories contribute to this pattern (for example, four species of *Bacillus* have ICLs and NADP-dependent IDHs and two species of *Rickettsia* have only NAD-dependent IDHs), they do not explain it. Members of both groups are highly diverse and include archaea, bacilli, and α- and γ-proteobacteria. Moreover, within each group are species with varied metabolic life-styles from diverse habitats. Each contains auto-, hetero-, chemo-, and lithotrophs; intra- and extracellular parasites; and organisms isolated from environments that vary from aerobic to anaerobic, from temperate to extremely hot, and from acidic to basic (*33*). Because the tight association between ICL and use of NADP by IDH is independent of taxonomic group, metabolic life-style, and physical habitat, we argue that the benefit to IDH of using NADP during growth on acetate is general and not particular to our experiments.

Conclusions. Was the switch in specificity from NAD to NADP by IDH adaptive or due to chance? The finding that the use of NAD arose on at least three independent occasions argues that the switch was adaptive. What was the basis of that adaptive switch? Physiological analysis using an *E. coli* IDH engineered to use NAD shows that use of NADP is strongly favored during growth on acetate. The 4% increase in growth rate in batch culture is a huge selective advantage that could produce a >50-fold increase in the ratio of icd^{NADP} to icd^{NAD} within just 100 generations. The 9% increase seen when *E. coli* strains were starved in chemostats could produce a >8000-fold enrichment within the same number of generations. Because all 46 prokaryotes with the ICL essential for growth on acetate also have an NADP-dependent IDH, and all 12 prokaryotes with only an NAD-dependent IDH have no ICL, the selection on acetate is not a laboratory artifact. Instead, it represents an adaptation to acetate use; one that was very likely the basis for the ancient recurrent switch to use of NADP by IDH.

We have shown that it is possible to reconstruct not only what occurred (the switch to NADP use by IDH) in an ancient adaptive event and how it occurred (changes at five specific amino acids in IDH), but also why it occurred (adaptation to growth on acetate or other similar energy-poor resources). Our studies provide a general approach to the reconstruction of ancient adaptive events.

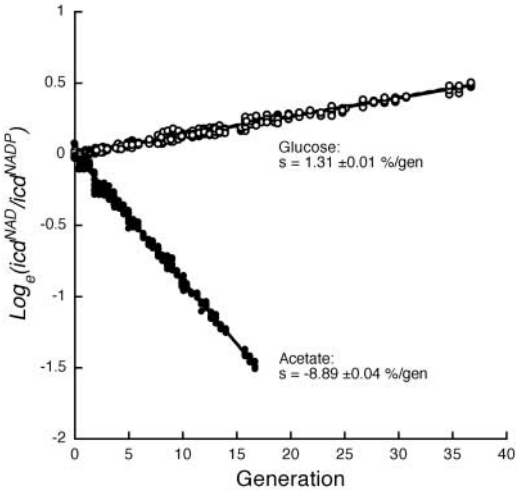


Fig. 3. Chemostat competition experiments. The slopes of the lines, *s*, are estimated selection coefficients per generation. The engineered icd^{NAD} is much less fit than the wild-type icd^{NADP} on acetate (solid circle), yet fitter on glucose (open circles). Presented are the pooled results from three chemostat competitions on each carbon source.

Table 1. Growth rates of icd^{NAD} relative to icd^{NADP} in various genetic backgrounds.

Genetic background	Relative growth rate (icd^{NAD}/icd^{NADP})*	
	Acetate	Glucose
Wild type	0.96 ± 0.02	0.998 ± 0.008
Δ*pntAB*	0.73 ± 0.04	0.991 ± 0.007
Δ*maeB*	0.63 ± 0.04	0.991 ± 0.006
Δ*pntAB*,Δ*maeB*	0.44 ± 0.03	0.963 ± 0.008
Δ*udhA*	1.46 ± 0.03	0.990 ± 0.004
Δ*udhA*,Δ*pntAB*	1.65 ± 0.09	0.992 ± 0.005
Δ*udhA*,Δ*maeB*	0.95 ± 0.11	0.975 ± 0.003
Δ*udhA*,Δ*pntAB*,Δ*maeB*	0.16 ± 0.05	0.974 ± 0.003

*Growth rates were determined in minimal medium with either acetate or glucose as the sole source of carbon and energy.

Phylogenetics provides molecular history. Structural biology reveals key replacements that are responsible for functional changes. Protein engineering tests the relationships between form and function. Studies of metabolism (physiology, development, and behavior in higher organisms) identify selectable phenotypes. Competition studies provide a means to explore the adaptive basis of any claim. Genomic comparisons confirm the generality of the results. By completing each of these steps, it is possible to identify the most probable cause of an ancient adaptive event that occurred billions of years ago.

References and Notes

1. G. B. Golding, A. M. Dean, *Mol. Biol. Evol.* **15**, 355 (1998).
2. T. M. Jermann, J. G. Opitz, J. Stackhouse, S. A. Benner, *Nature* **374**, 57 (1995).
3. J. Zhang, Y. P. Zhang, H. F. Rosenberg, *Nature Genet.* **30**, 411 (2002).
4. E. A. Gaucher, J. M. Thomson, M. F. Burgan, S. A. Benner, *Nature* **425**, 285 (2003).
5. Y. Shi, S. Yokoyama, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 8308 (2003).
6. B. Clarke, *Genetics* **79** (suppl.), 101 (1975).
7. D. Voet, J. G. Voet, *Biochemistry* (Wiley, New York, 1995).
8. P. A. Tipton, B. S. Beecher, *Arch. Biochem. Biophys.* **313**, 15 (1994).
9. H. Kirino, T. Oshima, *J. Biochem.* **109**, 852 (1991).
10. J. Miyazaki, N. Kobashi, M. Nishiyama, H. Yamane, *J. Biol. Chem.* **278**, 1864 (2003).
11. J. Sivaraman *et al.*, *J. Biol. Chem.* **278**, 43682 (2003).
12. J. Felsenstein, PHYLIP ver. 3.5 (Univ. of Washington, Seattle, WA, 1994).
13. A. M. Dean, G. B. Golding, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 3104 (1997).
14. Y. Yasutake *et al.*, *J. Biol. Chem.* **278**, 36897 (2003).
15. J. H. Hurley, A. M. Dean, D. E. Koshland Jr., R. M. Stroud, *Biochemistry* **30**, 8671 (1991).
16. J. H. Hurley, A. M. Dean, *Structure* **2**, 1007 (1994).
17. R. Chen, A. Greer, A. M. Dean, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 11666 (1995).
18. J. H. Hurley, R. Chen, A. M. Dean, *Biochemistry* **35**, 5670 (1996).
19. F. C. Neidhardt, J. L. Ingraham, M. Schaechter, *Physiology of the Bacterial Cell: A Molecular Approach* (Sinauer, Sunderland, MA, 1990).
20. D. M. Clarke, T. W. Loo, S. Gillam, P. D. Bragg, *Eur. J. Biochem.* **158**, 647 (1986).
21. J. Zhao, K. Shimizu, *J. Biotechnol.* **101**, 101 (2003).
22. P. E. Thorsness, D. E. Koshland Jr., *J. Biol. Chem.* **262**, 10422 (1987).
23. K. Walsh, D. E. Koshland Jr., *J. Biol. Chem.* **259**, 9646 (1984).
24. K. Walsh, D. E. Koshland Jr., *J. Biol. Chem.* **260**, 8430 (1985).
25. U. Sauer, F. Canonaco, S. Heri, A. Perrenoud, E. Fischer, *J. Biol. Chem.* **279**, 6613 (2004).
26. Materials and methods are available as supporting material on *Science* Online.
27. D. E. Dykhuizen, *Methods Enzymol.* **224**, 613 (1993).
28. M. Lunzer, A. Natarajan, D. E. Dykhuizen, A. M. Dean, *Genetics* **162**, 485 (2002).
29. A. M. Suiter, O. Bänzinger, A. M. Dean, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12782 (2003).
30. S. P. Miller *et al.*, *J. Biol. Chem.* **275**, 833 (2000).
31. K. A. Datsenko, B. L. Wanner, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6640 (2000).
32. G. Zhu, G. B. Golding, A. M. Dean, unpublished data.
33. M. T. Madigan, J. M. Martinko, J. Parker, *Brock Biology of Microorganisms* (Prentice-Hall, Upper Saddle River, NJ, ed. 9, 2000).

34. We thank M. Lunzer and S. Miller for technical assistance and B. Hall, B. Kerr, L. Merlo, S. Miller, and R. Redfield for constructive criticism of the manuscript. Supported by NIH grant GM060611 (A.M.D.).

1 November 2004; accepted 14 December 2004
Published online 13 January 2005;
10.1126/science.1106974
Include this information when citing this paper.

Axonopathy and Transport Deficits Early in the Pathogenesis of Alzheimer's Disease

Gorazd B. Stokin,[1] Concepción Lillo,[2] Tomás L. Falzone,[1] Richard G. Brusch,[1] Edward Rockenstein,[3] Stephanie L. Mount,[1] Rema Raman,[5] Peter Davies,[6] Eliezer Masliah,[3,4] David S. Williams,[2,3] Lawrence S. B. Goldstein[1]*

We identified axonal defects in mouse models of Alzheimer's disease that preceded known disease-related pathology by more than a year; we observed similar axonal defects in the early stages of Alzheimer's disease in humans. Axonal defects consisted of swellings that accumulated abnormal amounts of microtubule-associated and molecular motor proteins, organelles, and vesicles. Impairing axonal transport by reducing the dosage of a kinesin molecular motor protein enhanced the frequency of axonal defects and increased amyloid-β peptide levels and amyloid deposition. Reductions in microtubule-dependent transport may stimulate proteolytic processing of β-amyloid precursor protein, resulting in the development of senile plaques and Alzheimer's disease.

Axons and axonal transport exhibit prominent defects in a wide variety of neurological diseases. These defects often manifest as axonal swellings or spheroids, which correspond to axonal enlargements and aberrant accumulations of axonal cargos and cytoskeletal proteins (*1*). Molecular motor proteins propel axonal cargoes to and from presynaptic terminals along microtubule tracks and are thus crucial to understanding the role of impaired axonal transport in the pathogenesis of neurological diseases. A number of observations suggest that axonal transport may fail during the progression of Alzheimer's disease (AD) (*2*). We tested for axonal defects that are diagnostic of transport deficits and that might be related to early stages in the pathogenesis of AD.

AD is a common neurodegenerative disorder characterized by progressive cognitive deterioration and severe synaptic and neuronal loss. Pathological hallmarks of this disease—neurofibrillary tangles, neuropil threads, and senile plaques—are potentially linked to alterations of the axonal compartment (*3*). Neurofibrillary tangles and neuropil threads are related to the abnormal phosphorylation of the microtubule-associated protein tau and its dislocation from axons to presynaptic terminals and somatodendritic compartments. Senile plaques are composed of dystrophic neurites, some but not all of which are embedded in a matrix of extracellular amyloid. Some dystrophic neurites correspond to axonal swellings, which often contain abnormal accumulations of axonal cargos and tau (*4*).

Familial AD (FAD) mutations in β-amyloid precursor protein (βAPP) and in presenilin genes (PS1 and PS2) alter the production of amyloid-β peptides (Aβs), the major constituents of senile plaques, suggesting that proteolytic processing of βAPP into Aβs plays a central role in AD pathogenesis (*5*). In neurons, βAPP and its proteolytic machinery, PS1 and β-site βAPP cleaving enzyme (BACE), undergo kinesin-I–mediated fast anterograde axonal transport (*6*–*9*), during which βAPP may undergo proteolysis into Aβ-bearing intermediates (*10*) and Aβs (*7*, *11*). Increased Aβ levels (*11*, *12*) and senile plaques (*13*, *14*) in the axon-enriched white matter of mouse model and human AD brains suggest aberrant Aβ generation, or degradation, in axons. Reduced brain white matter in βAPP-deficient mice (*15*) also indicates that βAPP plays a role in axonal structure and function. Similar phenotypes have been found in a mouse model of AD (*16*) and in human AD (*17*), suggesting that axonal failure is an important part of AD. Overexpression of βAPP in *Drosophila* causes axonal transport defects, which are markedly enhanced by ordinarily

[1]Howard Hughes Medical Institute and Department of Cellular and Molecular Medicine, [2]Department of Pharmacology, [3]Department of Neurosciences, [4]Department of Pathology, [5]Department of Family and Preventive Medicine, School of Medicine, University of California San Diego (UCSD), 9500 Gilman Drive, La Jolla, CA 92093, USA. [6]Department of Pathology, F526, Albert Einstein College of Medicine, 1300 Morris Park Avenue, New York, NY 10461, USA.

*To whom correspondence should be addressed. E-mail: lgoldstein@ucsd.edu

benign reductions in the amount of kinesin molecular motor proteins (*18*). Finally, βAPP and Aβs accumulate in axonal swellings that resemble dystrophic neurites of senile plaques but instead form during brain aging (*19*) or after traumatic brain injury (*20*). These observations suggest a link between axonal transport of βAPP, aberrant Aβ generation, and AD pathology but do not untangle cause and effect relationships.

Early axonal defects in mouse AD models and human AD. To test whether FAD mutations of βAPP cause axonal defects, we analyzed a well-studied mouse AD model, Tg-swAPP^{Prp} (*21*), in which aberrant production of human Aβs results in deposition of senile plaques (*22*, *23*). Neuronal loss in the nucleus basalis of Meynert (NBM) is an invariable feature of AD (*24*), and alterations of cortical NBM projections are documented in aged primates (*25*), mouse models of AD (*26*, *27*), and human AD (*28*). We therefore examined fibers of the NBM, which provide the major cholinergic input to the cerebral cortex.

The organization of the NBM revealed by choline acetyltransferase (ChAT) staining was indistinguishable between 4- and 20-month-old wild-type and Tg-swAPP^{Prp} littermates. Unexpectedly, axonal varicosities in Tg-swAPP^{Prp} exhibited substantial variation in size and morphology. Generally, varicosities correspond to en passant synaptic boutons, are regularly spaced, and have relatively constant diameters. In contrast, varicosities in Tg-swAPP^{Prp} mice were often unusually large and irregularly spaced (Fig. 1A). Some displayed bizarre morphologies or lacked staining in their center, leaving only a ChAT-immunoreactive (IR) silhouette of the varicosity (Fig. 1B). Similar varicosities were occasionally encountered in the NBM of aged Tg-swAPP^{Prp} mice stained for phosphorylated tau (phospho-tau). To assess the microtubule-based transport machinery, we stained ChAT-IR fibers for phosphorylated high-molecular-weight neurofilament protein (phospho-NF-H) and kinesin light-chain subunits of kinesin-I (KLC) (Fig. 1C). Unlike varicosities, colocalization between ChAT, phospho-NF-H, and KLC in axonal swellings in Tg-swAPP^{Prp} was almost complete and very intense.

We quantified NBM ChAT-IR fibers by scoring coded samples for varicosities with diameters significantly larger than average in 4- and 20-month-old wild-type and Tg-swAPP^{Prp} mice (*21*). ChAT-IR varicosities of diameters larger than 3.0 μm were found almost exclusively in Tg-swAPP^{Prp} mice (Fig. 1D) and were paralleled by significant decreases in ChAT-IR fiber density in the NBM in 20-month-old Tg-swAPP^{Prp} mice compared with the wild type (Fig. 1E). To test whether abnormal varicosities exist in other mouse AD models, we examined Tg-sw/lonAPP^{Thy1} mice and found a significant increase in ChAT-IR varicosities of diameters larger than 3.0 μm in 2.5-month-old Tg-sw/lonAPP^{Thy1} mice compared with wild-type littermates (fig. S1).

To explore whether comparable swellings existed in other brain regions relevant to AD, we scored coded (*21*) Bielschowsky's silver-stained brain sections from 19- to 21-month-old wild-type and Tg-swAPP^{Prp} mice. Cortices and hippocampi of Tg-swAPP^{Prp}, but not wild-type, mice harbored frequent large swellings that resembled the ChAT-IR swellings in the NBM (Fig. 1F). Although morphologically similar to dystrophic neurites, ~50% of all swellings were found in brain areas devoid of amyloid (fig. S2).

To test for similar defects in AD, we probed for changes in ChAT- and phospho-tau-IR fibers in the human NBM. Brains were grouped for analysis into Braak AD stages 0, I–III, and IV–VI. We found a subset of ChAT- and phospho-tau-immunoreactive

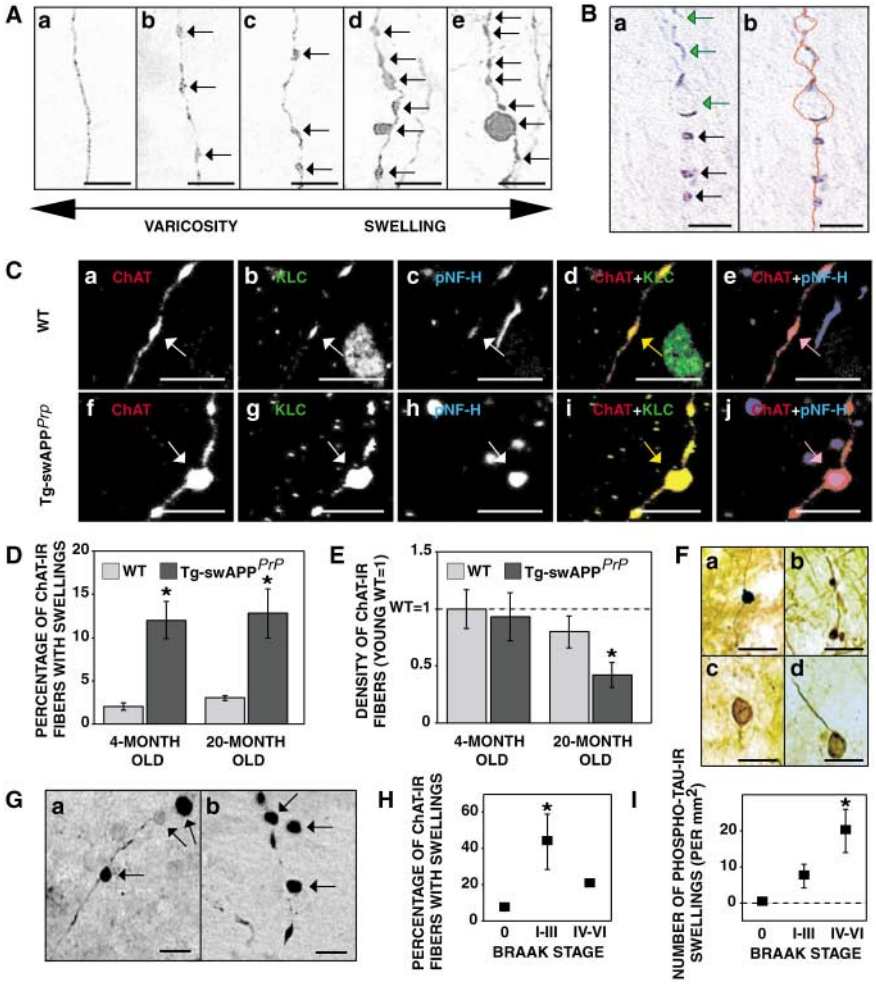


Fig. 1. Identification of axonal defects in mouse models and in human AD. (**A**) ChAT-IR fibers [(a) to (e)] with varicosities (arrows) of unusual size and morphology [(d) and (e)] in Tg-swAPP^{Prp} mouse NBM (scale bar, 10 μm). (**B**) Some varicosities exhibited ChAT-IR only within their perimeter, giving them a "ghost" appearance (normal varicosities, black arrows; "ghost" varicosities, green arrows; perimeter of "ghost" viscosities in (a) is traced in red in (b); scale bar, 10 μm). (**C**) Projections of stacks of optical sections of varicosities [(a) to (e)] and swellings [(f) to (j)] in the NBM, demonstrating accumulation of phospho-NF-H (pNF-H) and KLC in ChAT-IR swellings (scale bar, 10 μm). WT, wild-type. (**D**) Increased number of ChAT-IR fibers with varicosities of diameter larger than 3.0 μm (swellings) in the NBM of 4- ($P < 0.005$) and 20- ($P < 0.05$) month-old Tg-swAPP^{Prp} mice (100 fibers per NBM, $n = 4$ and 5, respectively), compared with wild-type mice (100 fibers per NBM, $n = 4$ and 5, respectively). (**E**) Density of ChAT-IR fibers is reduced in the NBM of 20-month-old Tg-swAPP^{Prp} ($n = 5$) but not wild-type ($n = 5$) mice ($P < 0.05$). (**F**) Swellings [(a) and (b)] and spheroids [(c) and (d)] in the cortex and hippocampus of 19- to 21-month-old Tg-swAPP^{Prp} mice ($n = 6$) stained with Bielschowsky's silver (scale bar, 10 μm). (**G**) ChAT- (a) and phospho-tau-IR (b) swellings (arrows) in the NBM of Braak II human AD (scale bar, 10 μm). (**H**) Increased number of ChAT-IR fibers with swellings in NBM in Braak group I–III ($n = 3$, $P < 0.05$) but not Braak group IV–VI ($n = 3$), compared with Braak group 0 ($n = 3$) human brains. Some error bars are too small to be visible. (**I**) Increased number of phospho-tau-IR swellings in NBM in Braak groups I–III ($n = 6$, $P = 0.07$) and IV–VI ($n = 3$, $P < 0.05$) compared with Braak group 0 ($n = 3$) human brains.

(phospho-tau-IR) fibers with swellings of morphology and diameters resembling those observed in the mouse model of AD (Fig. 1G). Differences in ChAT- [Krustal-Wallis (K-W) test, $P = 0.021$] (*21*) and phospho-tau-IR (K-W test, $P = 0.00021$) swellings between the Braak groups were significant. Compared to Braak group 0 (Fig. 1H), ChAT-IR swellings were common in Braak group I–III (up to 40 to 50% of all ChAT-IR fibers) but not in Braak group IV–VI. In contrast, the increase in phospho-tau-IR swellings observed in Braak group I–III became significant only in Braak group IV–VI, compared with Braak group 0 (Fig. 1I). The density of phospho-tau-IR fibers was also significantly increased in Braak group IV–VI but not in Braak group I–III, compared with Braak group 0 (fig. S3). Thus, axonal swellings occurred long before detectable amyloid deposition and, unexpectedly, accumulated excessive amounts of neurofilaments and kinesin-I. Formation of ChAT-IR swellings may represent an early event in AD pathogenesis, preceding or triggering the onset of known pathological changes such as aberrant tau phosphorylation.

Axonal swellings contain abnormal accumulations of organelles and vesicles. To evaluate the contents of axonal swellings, we stained NBM samples for ChAT and processed them for electron microscopy (Fig. 2A). We observed 5 to 7 ChAT-IR swollen profiles per section from 4-month-old Tg-swAPP^{Prp} mice and none from wild-type mice. Swellings contained large numbers of organelles and vesicles, were not myelinated, were not associated with postsynaptic densities, and frequently exhibited diameters larger than 3 μm (Fig. 2B). Similar profiles devoid of ChAT-IR were also observed (Fig. 2C). Tg-swAPP^{Prp} mice averaged 8 to 10 such profiles per section, in contrast to wild-type mice in which 2 to 3 were found per section at most. Swellings consisted of haphazardly arranged vesicles, mitochondria, sporadic multilamellar bodies, and vacuoles. Some were entirely filled by mitochondria, and most contained dense bodies that resembled those found in axons in early stages of Wallerian degeneration (*29*) and in dystrophic neurites embedded in amyloid in AD (*30*). A subset of swellings encountered exclusively in the Tg-swAPP^{Prp} mice displayed characteristics of axonal degeneration, including electron-dense granular axoplasm and a considerable amount of axoplasmic debris of tubular appearance (Fig. 2D).

Formation of axonal swellings is enhanced by reduction in kinesin-I. Aberrant accumulation of kinesin-I, organelles, and vesicles within axonal swellings is suggestive of impaired axonal transport. To test whether impairing anterograde axonal transport enhances the formation of axonal swellings, as is seen in *Drosophila* models of AD (*18*, *31*), we reduced the genetic dosage of KLC1 by 50% in Tg-swAPP^{Prp} mice (figs. S4 and S5). Axonal swellings that were independent of amyloid deposits exhibited a significant increase in average number per brain in Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} compared to Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} mice (Fig. 3, A and B). This increase appeared region-specific and was most prominent in select cortical regions (Fig. 3C). There was no significant difference in the average number of axonal swellings per amyloid deposit between Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} and Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} mice (fig. S6).

We directly tested and confirmed that a relatively benign reduction in kinesin-I is sufficient to impair axonal transport of βAPP. We transiently transfected neuronal cultures harvested from KLC1^{wt}/KLC1^{wt} and KLC1^{wt}/KLC1^{null} mouse hippocampi with wild-type βAPP linked to yellow fluorescent protein (APPYFP). In KLC1^{wt}/KLC1^{null}, we observed a significant reduction in the percentage of APPYFP particles undergoing anterograde transport and an increase in the percentage of APPYFP particles undergoing retrograde transport compared with KLC1^{wt}/KLC1^{wt} primary hippocampal cultures (Fig. 3, D and E, and movies S1 and S2). There was no difference in the percentage of

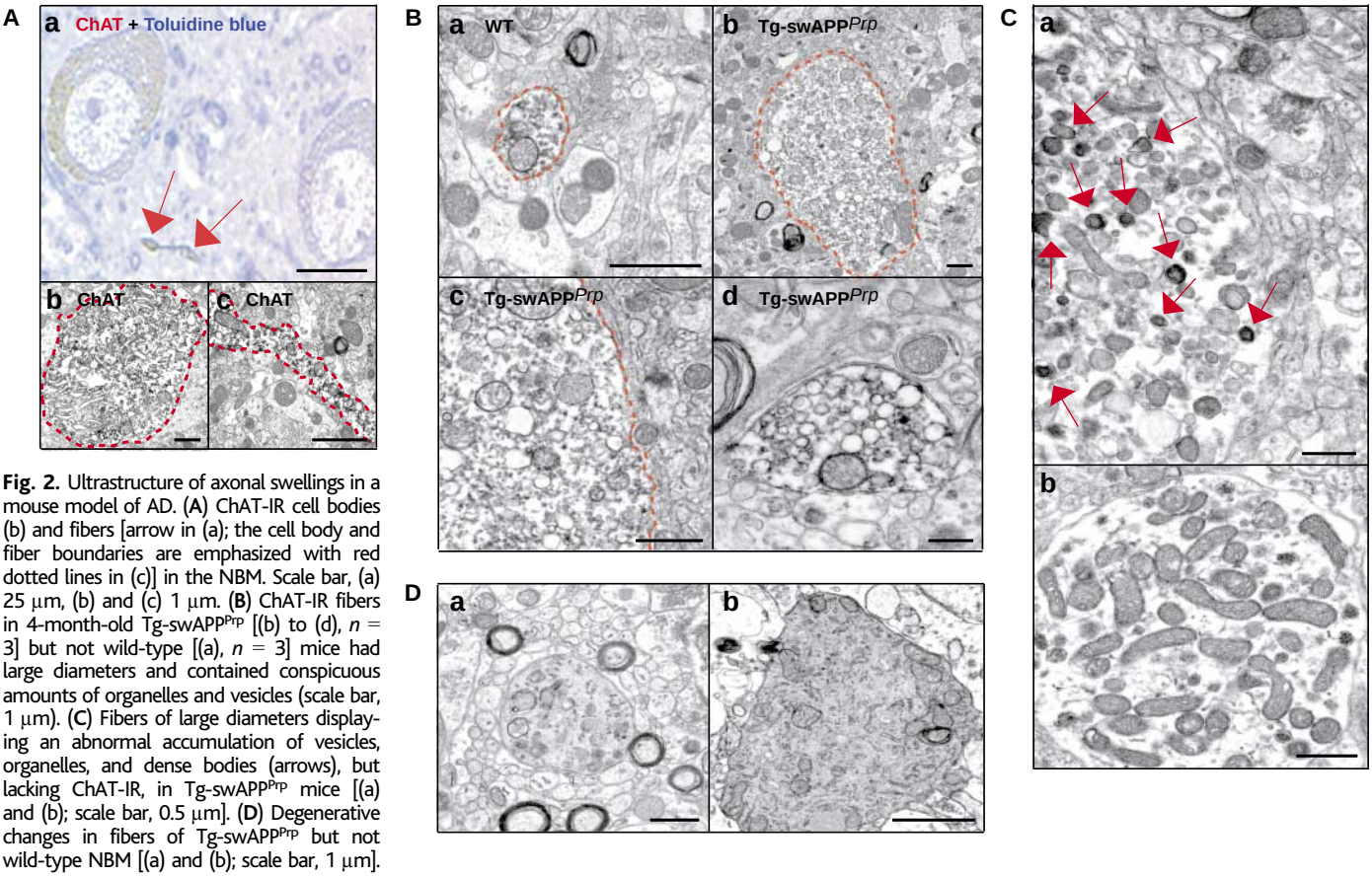


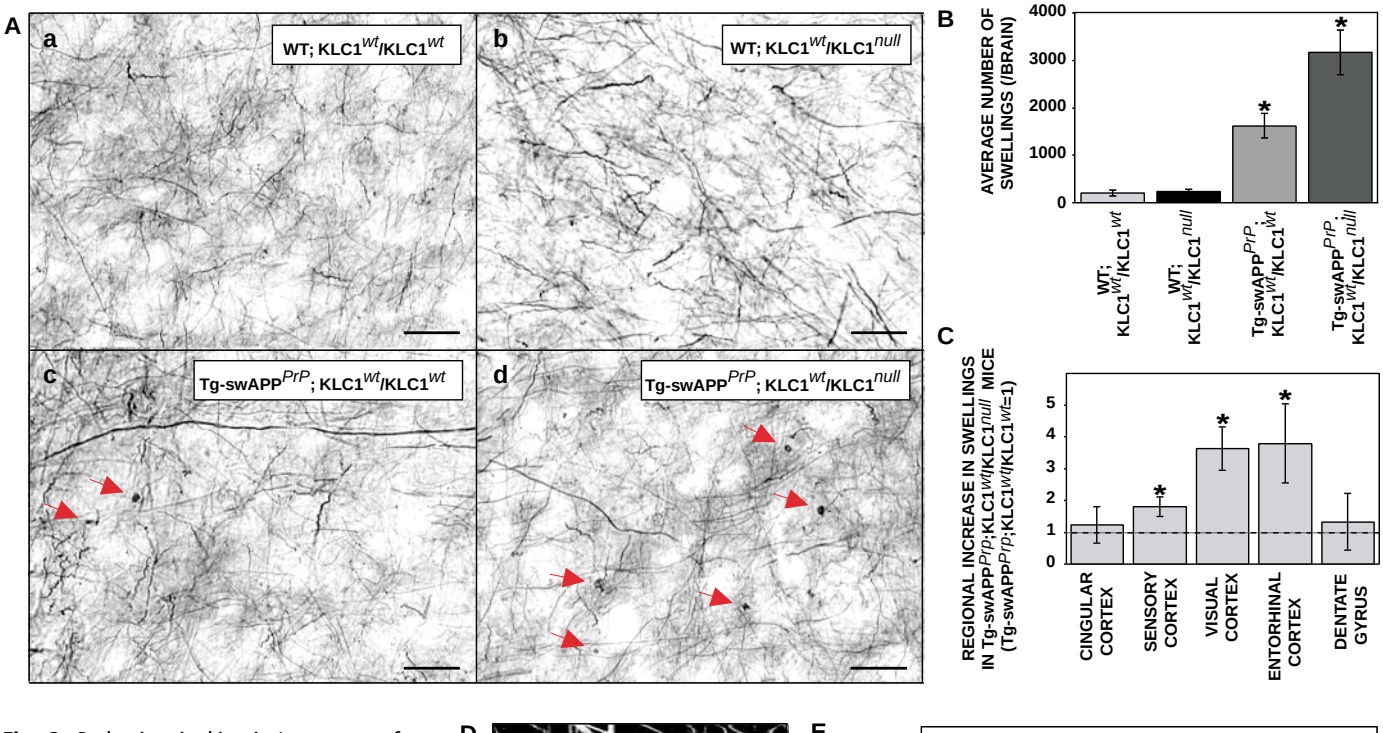
Fig. 2. Ultrastructure of axonal swellings in a mouse model of AD. (**A**) ChAT-IR cell bodies (b) and fibers [arrow in (a); the cell body and fiber boundaries are emphasized with red dotted lines in (c)] in the NBM. Scale bar, (a) 25 μm, (b) and (c) 1 μm. (**B**) ChAT-IR fibers in 4-month-old Tg-swAPP^{Prp} [(b) to (d), $n = 3$] but not wild-type [(a), $n = 3$] mice had large diameters and contained conspicuous amounts of organelles and vesicles (scale bar, 1 μm). (**C**) Fibers of large diameters displaying an abnormal accumulation of vesicles, organelles, and dense bodies (arrows), but lacking ChAT-IR, in Tg-swAPP^{Prp} mice [(a) and (b); scale bar, 0.5 μm]. (**D**) Degenerative changes in fibers of Tg-swAPP^{Prp} but not wild-type NBM [(a) and (b); scale bar, 1 μm].

stalled APPYFP particles. Similar results were obtained in *Drosophila* (figs. S7 and S8 and movies S3 to S5). Thus, a phenotypically benign 50% reduction in kinesin-I subunits was sufficient to produce marked impairment in anterograde transport of βAPP accompanied by enhanced formation of axonal swellings, independently of amyloid deposition.

Reduction in kinesin-I increases Aβ generation and its intraneuronal accumulation. Although full-length βAPP levels were indistinguishable between Tg-swAPP^{Prp};KLC1^{wt}/KLC^{wt} and Tg-swAPP^{Prp};KLC1^{wt}/KLC^{null} mice (Fig. 4, A and B), levels of Aβs as well as Aβ42/Aβ40 ratios were consistently elevated in Tg-swAPP^{Prp};KLC1^{wt}/KLC^{null} mice (Fig. 4, C and D). These increases in the Aβ42/Aβ40 ratio were not a result of genetic background effects and were comparable in magnitude to well-established Aβ42/Aβ40 ratio increases caused by an FAD mutant PS1 transgene (fig. S9). General effects of axonal structure and microtubule organization on Aβ42/Aβ40 ratios were ruled out by analyzing a deletion of the low-molecular-weight neurofilament subunit (NF-L) (fig. S9).

Aβs may accumulate intracellularly in AD brains, possibly before extracellular amyloid deposition (*32*). We asked if this feature was present in Tg-swAPP^{Prp} and whether it could be enhanced by KLC1 reduction. Antibodies against the C terminus of βAPP and against the N terminus of the Aβ sequence of βAPP primarily stained the somatodendritic compartment throughout the brain, with the latter showing a discrete staining of the corpus callosum (fig. S10). The staining pattern with an antibody against the C terminus of Aβ40 corresponded to that observed for βAPP and was indistinguishable between 17- to 22-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} mice and Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} mice. However, an antibody directed against the C terminus of Aβ42 exhibited significantly increased staining within cell bodies in the sensory and entorhinal cortices, but not in the hippocampi or dentate gyri, of 17- to 22-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} mice compared with Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} mice (Fig. 5, A and B). Evidence that KLC1 reduction caused accumulation of Aβ42 in axons came from the finding that human Aβ42 and Aβ42/Aβ40 ratios were increased in the sciatic nerves of 11- to 17-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} compared with Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} mice. Aβ42/Aβ40 ratios were also significantly higher in the sciatic nerves than in the brains of Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} but not Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} mice (Fig. 5C). Thus, Aβ42 formation could be enhanced by the reduction of KLC1, such that it accumulates intraneuronally in Tg-swAPP^{Prp} mice at an age that precedes or coincides with the onset of amyloid deposition (*23*).

Reduction of kinesin-I enhances amyloid deposition. To determine whether the



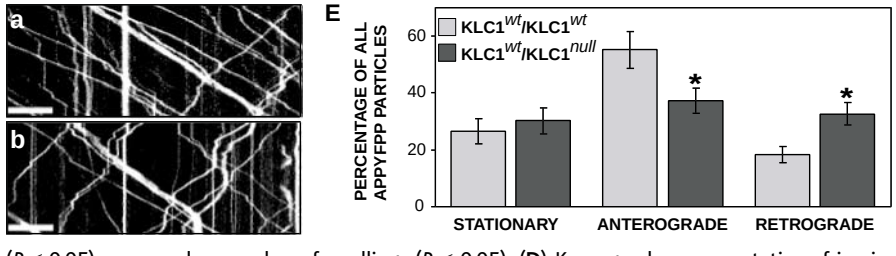


Fig. 3. Reduction in kinesin-I promotes formation of axonal swellings. (**A**) Swellings and spheroids (arrows) in Bielschowsky's silver stained sensory cortices of 19-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} (c) and Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} (d) but not WT;KLC1^{wt}/KLC1^{wt} (a) or WT;KLC1^{wt}/KLC1^{null} (b) mice (scale bar, 25 μm). (**B**) Quantification of axonal defects. Nineteen- to 21-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} mice (n = 6) had significantly increased numbers of swellings per brain ($P < 0.05$) compared with Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} mice (n = 7). Swellings in 19- to 21-month-old WT;KLC1^{wt}/KLC1^{wt} (n = 4) or WT;KLC1^{wt}/KLC1^{null} (n = 3) brains were rare. (**C**) Increased numbers of swellings upon deletion of one copy of KLC1 was region specific: Sensory, visual, and entorhinal cortex, but not cingulate cortex or dentate gyrus, had significant increases in the number of swellings ($P < 0.05$). (**D**) Kymographs representative of in vivo transport of APPYFP in KLC1^{wt}/KLC1^{wt} [(a), n = 9] and KLC1^{wt}/KLC^{null} [(b), n = 14] hippocampal cultures (scale bar, 10 μm). (**E**) Altered percentage of APPYFP particles traveling in anterograde and retrograde directions in KLC1^{wt}/KLC^{null} (n = 292, $P < 0.005$) compared to KLC1^{wt}/KLC1^{wt} (n = 193, $P < 0.005$) hippocampal cultures.

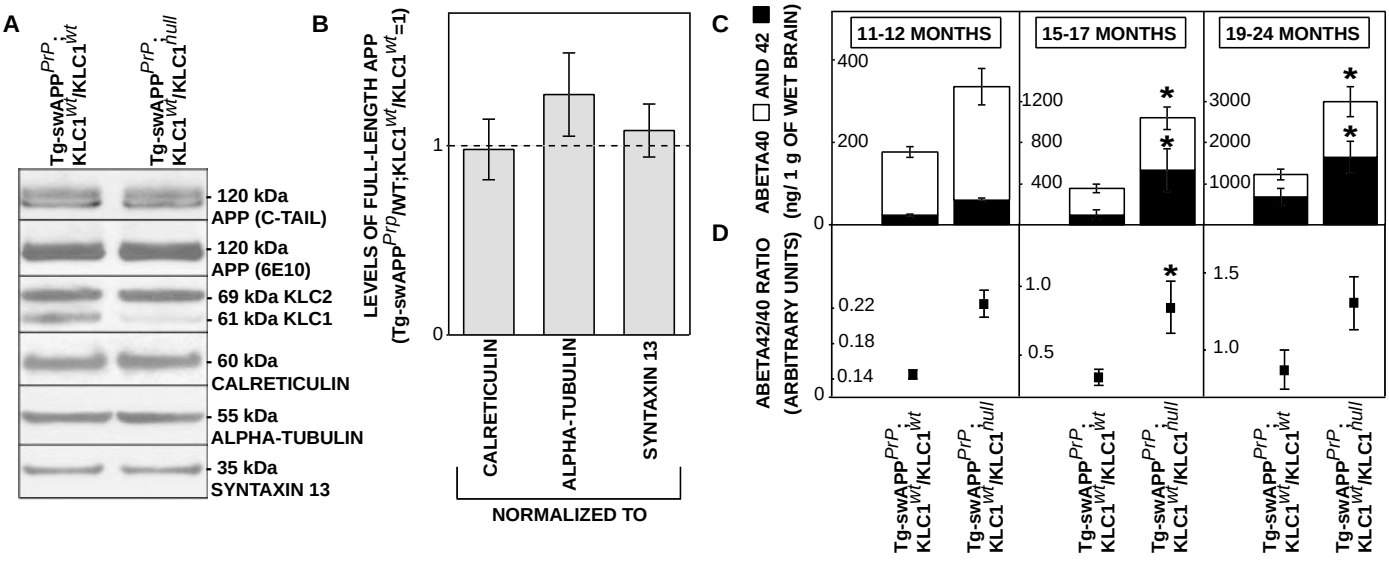


Fig. 4. Reduction in kinesin-I increases levels of brain Aβs. (**A**) Immunolabeled Western blot showing that full-length βAPP levels were similar, whereas KLC1 was specifically reduced in 15- to 17-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC^{null} (n = 7) compared to Tg-swAPP^{Prp};KLC1^{wt}/KLC^{wt} (n = 7) mice. (**B**) Quantitation of full-length βAPP from Western blots as in (A) in 11- to 24-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC^{wt} (n = 15) and Tg-swAPP^{Prp};KLC1^{wt}/KLC^{null} (n = 17) mice. (**C**) Eleven- to 12-, 15- to 17-, and 19- to 24-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} mice (n = 2, 7, and 8, respectively) had increased levels of brain Aβs, determined by enzyme-linked immunosorbent assay (*21*), compared with Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} mice (n = 2, 7, and 6, respectively; $P < 0.05$). No Aβs were detected in WT;KLC1^{wt}/KLC^{wt} (n = 5) or WT;KLC1^{wt}/KLC^{null} (n = 3) mice. (**D**) Eleven- to 12- and 15- to 17-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC^{null} mice (n = 9) had increased Aβ42/Aβ40 ratios when compared with the Tg-swAPP^{Prp};KLC1^{wt}/KLC^{wt} (n = 9) mice ($P < 0.05$).

increased Aβ42/Aβ40 ratio observed upon reduction of KLC1 correlated with accelerated or enhanced amyloid deposition, we examined senile plaques in 18- to 24-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC^{wt} and Tg-swAPP^{Prp};KLC1^{wt}/KLC^{null} mice (*21*) (Fig. 6A). The average number of amyloid plaques visualized with thioflavine S and Bielschowsky's silver was significantly increased in Tg-swAPP^{Prp};KLC1^{wt}/KLC^{null} compared with Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} brains (Fig. 6, B and C). Average diameters and volumes of Bielschowsky's silver-stained senile plaques were also significantly increased in Tg-swAPP^{Prp};KLC1^{wt}/KLC^{null} compared with the Tg-swAPP^{Prp};KLC1^{wt}/KLC^{wt} mice (fig. S11). The most marked increase in the number of plaques was visualized by staining for C termini of Aβ40 and Aβ42. Diffuse extracellular Aβ immunoreactivity was found almost exclusively in Tg-swAPP^{Prp};KLC1^{wt}/KLC^{null} brains (fig. S12), suggesting that reduction in KLC1 also promotes the formation of noncongophilic and nonfibrillar deposits. The average number of thioflavine-S–positive senile plaques in 11- to 12-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC^{null} mice was comparable to that in 17- to 19- or 20- to 24-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC^{wt} controls (Fig. 6D and fig. S13). This number increased in aged mice and far exceeded the number of amyloid deposits in older Tg-swAPP^{Prp};KLC1^{wt}/KLC^{wt} mice. Enhanced amyloid deposition in Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} mice was most prominent in the same cortical regions as the swellings not associated with amyloid (Fig. 6E). Moreover, a significant correlation between the number of swellings not associated with amyloid and the number of senile plaques was observed in all transgenic mice examined (fig. S14), suggesting that swellings precede and participate in the formation of amyloid plaques.

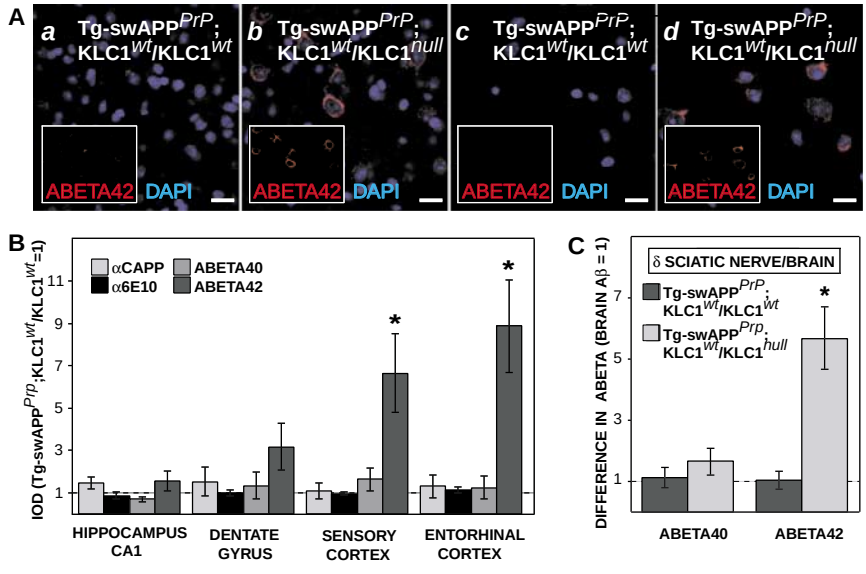


Fig. 5. Aberrant intraneuronal accumulation of Aβs upon reduction of kinesin-I. (**A**) Staining for Aβ42, counterstained with 4′,6-diamidino-2-phenylindole (DAPI), in 19-month-old sensory [(a) and (b)] and entorhinal [(c) and (d)] cortices of the Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} [(b) and (d)] and Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} [(a) and (c)] mice (scale bar, 25 μm). (**B**) Analysis of integrated optical densities (IOD) for βAPP and Aβ in the sensory cortex, entorhinal cortex, pyramidal layer of the CA1, and granular layer of the dentate showed that 17- to 22-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} (n = 5) mice had greater Aβ42-IOD values in sensory and entorhinal cortices than Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} (n = 5) controls ($P < 0.05$). (**C**) Increased sciatic nerve versus brain Aβ42/Aβ40 ratio in 11- to 17-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} (n = 3) compared with Tg-swAPP^{Prp};KLC1^{wt}/KLC^{wt} (n = 3) mice ($P < 0.05$).

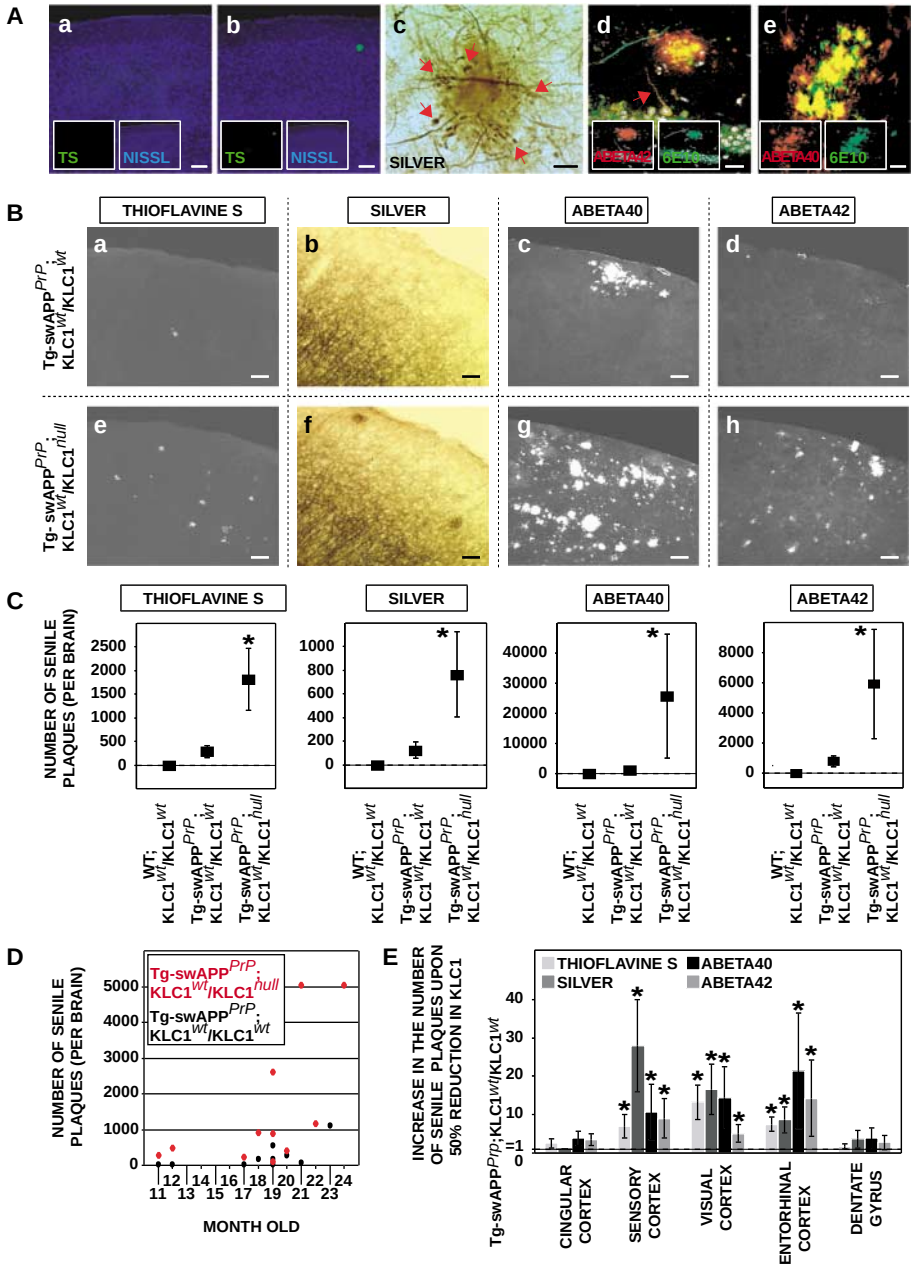


Fig. 6. Reduction in kinesin-I accelerates and enhances amyloid deposition. (**A**) Brain sections from Tg-swAPP^{Prp} (b), but not wild-type (a), mice showed thioflavine S–positive (TS) senile plaques (Nissl-stained visual cortex in blue; scale bar, 50 μm) surrounded by a corona of dystrophic neurites (arrows; scale bar, 25 μm) when stained with Bielschowsky's silver (c). Plaques were also stained for the C terminus of Aβ40 and Aβ42 or against its N terminus, 6E10 [(d) and (e); scale bar, 10 μm); Aβ40 immunoreactivity is visible in the projection (arrow) from the CA1 layer (d). (**B**) Enhanced senile plaque deposition in sensory cortices (scale bar, 100 μm) of 21-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} [(e) to (h)] compared with Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} [(a) to (d)] mice. (**C**) Increase in average number of senile plaques visualized by four independent stainings in 18- to 24-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} (n = 8) compared with Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} mice (n = 7, $P < 0.05$). (**D**) Significant overall difference ($P < 0.05$) in age of onset and progression of amyloid deposition in 11- to 24-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} (n = 10) and Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} (n = 13) mice assayed by thioflavine S. (**E**) Regional differences in amyloid deposition between 18- to 24-month-old Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{wt} (n = 7) and Tg-swAPP^{Prp};KLC1^{wt}/KLC1^{null} (n = 8, $P < 0.05$) mice.

Discussion. Our findings suggest that axonal blockages that interfere with axonal transport occur early in the course of AD and in mouse models of AD. Morphologically, these blockages resembled axonal swellings, a hallmark of axonal injury, and dystrophic neurites embedded in amyloid within senile plaques of AD (*4*, *30*). Unlike dystrophic neurites, which are primarily associated with amyloid deposits in AD (*33*), axonal swellings preceded amyloid and other disease-related pathology in a mouse AD model by at least a year and were found in brain areas lacking amyloid deposition in the early stages of human AD. Thus, axonal swellings do not form in response to amyloid deposition. Rather, axonal swellings may represent precursors of some form of dystrophic neurites. Reports of gross white matter changes in mice transgenic (*16*) or deficient (*15*) for βAPP and in human AD (*17*) are consistent with this proposal.

Axonal transport deficits were previously found during late stages of AD (*34*). Our data suggest that axonal transport deficits also play an early, and potentially causative, role in AD. For example, direct relevance to AD of axonal phenotypes caused by mutant βAPP overexpression in *Drosophila* and mice comes from human studies suggesting that overexpression of βAPP is responsible for the AD pathology observed in Down syndrome (*35*). Similarly, our samples of early-stage human AD brains exhibited swellings in cholinergic axons of the NBM akin to those observed in mouse AD models. The observation that reduction of kinesin-I dosage causes an increase in the proportion of βAPP undergoing retrograde transport directly correlates with the accumulation of Aβs in cell bodies and amyloid deposition in the vicinity of cell bodies of the entorhinal cortex within the perforant pathway. These findings also suggest that enhanced amyloidogenesis in the entorhinal cortex is not the result of amyloid deposition by afferents terminating in this region.

Together, the chronological and genetic relationships between reduced axonal transport, axonal swellings, Aβ generation, and amyloid deposition suggest strong mechanistic ties among these events. Axonal swellings could form because of impaired axonal transport and promote aberrant Aβ generation. If aberrant Aβ generation occurs locally at sites of blockage, then amyloid deposition may occur as a result of focally increased secretion of Aβs or lysis of Aβ-enriched axonal swellings. Either, or both, of these processes may provide an appropriate substrate for the formation of senile plaques. Indeed, accumulation of Aβs is observed at sites of axonal damage after traumatic brain injury in the central nervous system (*20*, *36*). Alternatively, upon encountering an axonal swelling or blockage, vesicles containing βAPP or Aβs might undergo retrograde transport to the somatodendritic compartment, where aberrant Aβ generation and/or amyloid deposition could occur. This process could be enhanced by genetic reduction of kinesin-I, which we found increased the probability that an APP vesicle entered the retrograde transport pathway. βAPP is also proposed to play a critical role in linking

kinesin-I to axonal vesicles (*6*, *7*). If βAPP processing to Aβs occurs at axonal (or dendritic) blockages and causes kinesin-I release from vesicles and reduction of axonal transport (*7*), then blockages may lead to local stimulation of βAPP processing, which in turn would cause additional vesicle stalling and further local stimulation of βAPP processing. This proposed sequence of events would generate an autocatalytic spiral in which processes leading to axonal blockages and Aβ production become mutually stimulatory. A vicious cycle of axonal blockages and aberrant Aβ generation also provides a rational explanation for early synaptic loss in AD. Such a mechanism would be critical in AD linked to polymorphisms in KLC1 (*37*) as well as in sporadic AD, which could potentially be initiated by likely differences in age-dependent declines in axonal transport among humans.

References and Notes

1. S. Yagashita, *Virchows Arch. A Pathol. Anat. Histol.* **378**, 181 (1978).
2. A. D. Cash *et al.*, *Am. J. Pathol.* **162**, 1623 (2003).
3. R. D. Terry, *J. Neuropathol. Exp. Neurol.* **55**, 1023 (1996).
4. E. Masliah *et al.*, *J. Neuropathol. Exp. Neurol.* **52**, 619 (1993).
5. D. J. Selkoe, *Nature* **399**, A23 (1999).
6. A. Kamal, G. B. Stokin, Z. Yang, C. H. Xia, L. S. Goldstein, *Neuron* **28**, 449 (2000).
7. A. Kamal, A. Almenar-Queralt, J. F. LeBlanc, E. A. Roberts, L. S. Goldstein, *Nature* **414**, 643 (2001).
8. H. Papp, M. Pakaski, P. Kasa, *Neurochem. Int.* **41**, 429 (2002).
9. J. G. Sheng, D. L. Price, V. E. Koliatsos, *Exp. Neurol.* **184**, 1053 (2003).
10. P. J. Morin *et al.*, *J. Neurochem.* **61**, 464 (1993).
11. A. E. Roher *et al.*, *Biochemistry* **41**, 11080 (2002).
12. O. Wirths *et al.*, *Brain Pathol.* **12**, 275 (2002).
13. T. Uchihara, H. Kondo, H. Akiyama, K. Ikeda, *Acta Neuropathol.* **90**, 51 (1995).
14. L. Holcomb *et al.*, *Nature Med.* **4**, 97 (1998).
15. U. Muller *et al.*, *Cell* **79**, 755 (1994).
16. J. M. Redwine *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 1381 (2003).
17. S. E. Rose *et al.*, *J. Neurol. Neurosurg. Psychiatry* **69**, 528 (2000).
18. S. Gunawardena, L. S. Goldstein, *Neuron* **32**, 389 (2001).
19. T. Kawarabayashi *et al.*, *Neurosci. Lett.* **153**, 73 (1993).
20. D. H. Smith, X. H. Chen, A. Iwata, D. I. Graham, *J. Neurosurg.* **98**, 1072 (2003).
21. Materials and methods are available as supporting material on *Science* Online.
22. D. R. Borchelt *et al.*, *Neuron* **17**, 1005 (1996).
23. D. R. Borchelt *et al.*, *Neuron* **19**, 939 (1997).
24. P. J. Whitehouse *et al.*, *Science* **215**, 1237 (1982).
25. C. A. Kitt *et al.*, *Science* **226**, 1443 (1984).
26. T. P. Wong, T. Debeir, K. Duff, A. C. Cuello, *J. Neurosci.* **19**, 2706 (1999).
27. D. Hernandez *et al.*, *Neuroreport* **12**, 1377 (2001).
28. P. Davies, A. J. Maloney, *Lancet* **2**, 1403 (1976).
29. H. D. Webster, *J. Cell Biol.* **12**, 361 (1962).
30. R. D. Terry, N. K. Gonatas, M. Weiss, *Am. J. Pathol.* **44**, 269 (1964).
31. L. Torroja, H. Chu, I. Kotovsky, K. White, *Curr. Biol.* **9**, 489 (1999).
32. G. K. Gouras *et al.*, *Am. J. Pathol.* **156**, 15 (2000).
33. E. Masliah *et al.*, *Am. J. Pathol.* **142**, 871 (1993).
34. A. Salehi, J. D. Delcroix, W. C. Mobley, *Trends Neurosci.* **26**, 73 (2003).
35. V. P. Prasher *et al.*, *Ann. Neurol.* **43**, 380 (1998).
36. X. H. Chen *et al.*, *Am. J. Pathol.* **165**, 357 (2004).
37. C. M. Dhaenens *et al.*, *Neurosci. Lett.* **368**, 290 (2004).

38. We thank D. W. Cleveland, F. H. Gage, P. R. Mouton, and R. D. Terry for helpful discussions; M. Yasuda for help with primary hippocampal cultures; M. P. Sundsmo and the UCSD Alzheimer's Disease Research Center (ADRC) for human samples; and the UCSD Cancer Center Digital Imaging Shared Resource for help with imaging. Supported by an Ellison Medical Foundation Senior Scholar Award in Aging Research (L.S.B.G.); NIH grant nos. EY13408, EY12598 (D.S.W.), and P50 AG05131 (A.D.R.C.); a Pew Foundation fellowship (T.L.F.); and a Boehringer-Ingelheim Fonds fellowship (G.B.S.). L.S.B.G. is an Investigator of the Howard Hughes Medical Institute.

Supporting Online Material

www.sciencemag.org/cgi/content/full/307/5713/1282/DC1
Materials and Methods
Figs. S1 to S14
References and Notes
Movies S1 to S5

27 September 2004; accepted 2 December 2004
10.1126/science.1105681

REPORTS

The Use of Transit Timing to Detect Terrestrial-Mass Extrasolar Planets

Matthew J. Holman[1]* and Norman W. Murray[2]

Future surveys for transiting extrasolar planets are expected to detect hundreds of jovian-mass planets and tens of terrestrial-mass planets. For many of these newly discovered planets, the intervals between successive transits will be measured with an accuracy of 0.1 to 100 minutes. We show that these timing measurements will allow for the detection of additional planets in the system (not necessarily transiting) by their gravitational interaction with the transiting planet. The transit-time variations depend on the mass of the additional planet, and in some cases terrestrial-mass planets will produce a measurable effect. In systems where two planets are seen to transit, the density of both planets can be determined without radial-velocity observations.

About 130 extrasolar planets (*1*) have been detected by (i) short-duration brightness anomalies in gravitational microlensing events caused by planets near the lens star (*2*), (ii) reflex motions of the central star (as revealed by radial-velocity variations in the stellar spectrum or radio pulse arrival times) (*3*–*6*), and (iii) variations in the apparent stellar brightness caused by a planetary transit (passage of the planet in front of the star) (*7*–*11*). These approaches provide complementary information. Gravitational microlensing measurements primarily constrain the ratio of planet mass to stellar mass. Radial-velocity measurements lead to estimates of the orbital period, eccentricity, and minimum mass of the planet. With present technology, radial-velocity surveys can only detect planets with masses greater than about 10 $M_{\odot}$ (orbiting low-mass stars) (*12*–*14*). Transit observations provide estimates of the orbital period and planetary radius. Transit surveys, particularly if space based, will be sensitive to planets as small as Mercury for the smallest radius stars observed (*15*). Barring major improvements in the precision of radial-velocity measurements, the measurement of the mass and radius (and thus the average density) of a terrestrial-sized extrasolar planet would appear to be out of reach.

Here, we point out that variations in the time interval between transits, produced by gravitational interactions with additional planets, allow for the orbital period and mass of the additional planet to be determined from transit observations alone.

The time interval between successive transits of an unperturbed planet is always the same [aside from small corrections due to orbital precession (*16*) or to the decay of the transiting planet's orbital semimajor axis as a result of tidal interaction with the star (*17*)]. However, for stellar systems, the presence of a third star orbiting a stellar binary can produce short-term variations, in addition to the more familiar long-term variations, of the period of the binary (*18*, *19*). The same is true for planetary systems. The interval between successive transits of the extrasolar planet HD

[1]Harvard-Smithsonian Center for Astrophysics, MS51, 60 Garden Street, Cambridge, MA 02138, USA. [2]Canadian Institute for Theoretical Astrophysics, University of Toronto, 60 St. George Street, Toronto, ON M5S 3H8, Canada.

*To whom correspondence should be addressed. E-mail: mholman@cfa.harvard.edu

209458b would vary by ±3 s if a second planet (of mass $10^{-4}\ M_{\odot}$, period 80 days, and eccentricity $e \sim 0.4$) existed in that system (*20*).

Over the course of their orbits, the transiting planet and a second planet exchange energy and angular momentum as a result of their mutual gravitational interaction. This interaction, greatest at each planetary conjunction, results in short-term oscillations of the semimajor axes and eccentricities of the planets, which in turn alter the interval between successive transits. We illustrate this effect by considering our solar system (Fig. 1). The gravitational perturbations among the planets in the solar system lead to transit-interval variations ranging from tens of seconds for Mercury to thousands of seconds for Mars. The variations for Earth and Venus show oscillations with the 583-day Earth-Venus synodic period. For an exterior transiting planet, the transit interval can be altered by an inner planet's indirect influence on the position of the star, even if the gravitational interaction between the planets is negligible (*21*).

Next, we investigate the influence of an additional Jupiter-mass planet on the transit interval of HD 209458b. We integrate the heliocentric equations of motion of the hypothetical two-planet system (*22*). We assume that the two planets are coplanar, have orbits that are perpendicular to the sky plane, and have initially aligned orbital apsides. We include the mutual gravitational interactions of the planets and terms that account for the general relativistic (GR) influence of the central star (*23*) but neglect the terms for the GR influence of the planetary masses, terms for the oblateness of the star, and terms for the tidal interaction with the star. During the course of each integration, we iteratively solve for the time at which the centers of star and planet reach their minimum projected separation.

Sharp increases in the transit interval occur near the times when the outer planet reaches its periastron (Fig. 2). Hence, the orbital periods of both planets in each of the examples can be determined directly. We chose the period and eccentricity of the hypothetical perturbing planets such that their pericenter distances are roughly the same. As a result, the largest transit-interval variations have comparable magnitude. For clarity, we ignore the light-time effect (*21*) (due to the varying distance between the star and the observer as the star moves with respect to the center of mass of the star and planets) from the dynamical effects. The magnitude of the light-time effect due to a Jupiter-mass perturber with a 1-AU (astronomical unit) semimajor axis orbiting a solar-mass star is 0.5 s.

Whether the presence of a companion planet can be detected from short-term transit interval variations depends on the difference between the minimum and maximum transit interval. We consider jovian-mass ($10^{-3}\ M_{\odot}$) and terrestrial-mass ($3 \times 10^{-6}\ M_{\odot}$) perturbers. A range of perturber periods is tested, in increments of 0.1 day, starting with the minimum perturber period that ensures that the orbits of the two planets do not initially cross. We follow the same numerical procedure described earlier. During each integration, simulating 10^4 days, we recorded the times of transit, from which the minimum and maximum intervals between transits were determined. For a given perturber eccentricity, the period variation decreases as the perturber period and semimajor axis increase. Likewise, for a given perturber period, the variation is greater for larger perturber eccentricity. The transit-interval variation is primarily a function of the periastron distance of the perturber, and a com-

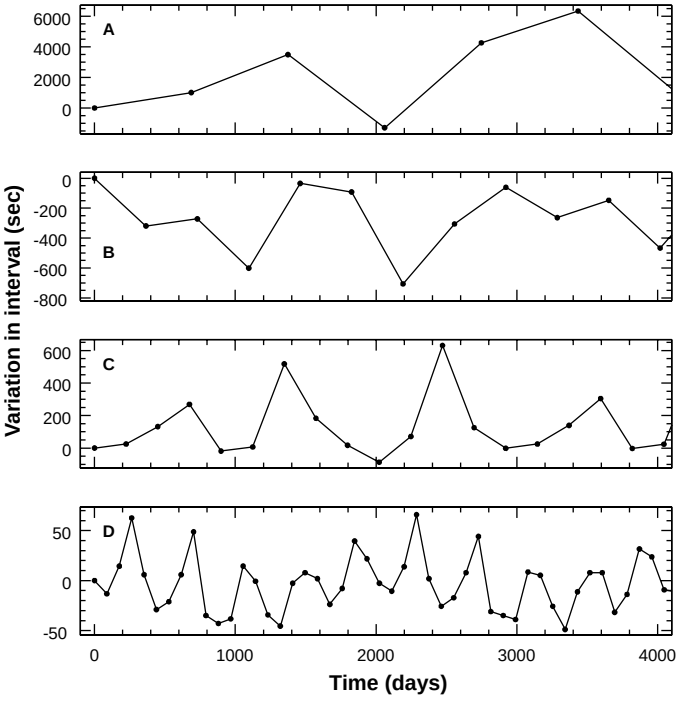


Fig. 1. The variations of the interval between successive transits of terrestrial planets, induced by the other planets in the solar system. (**A** to **D**) The variations for Mars, Earth, Venus, and Mercury, respectively. To guide the eye, the solid line connects the times of each transit. The transit intervals result from numerically integrating the equations of motion of the planets in our solar system and calculating the transit times as seen by distant observers located in the present-day orbital planes of the various planets.

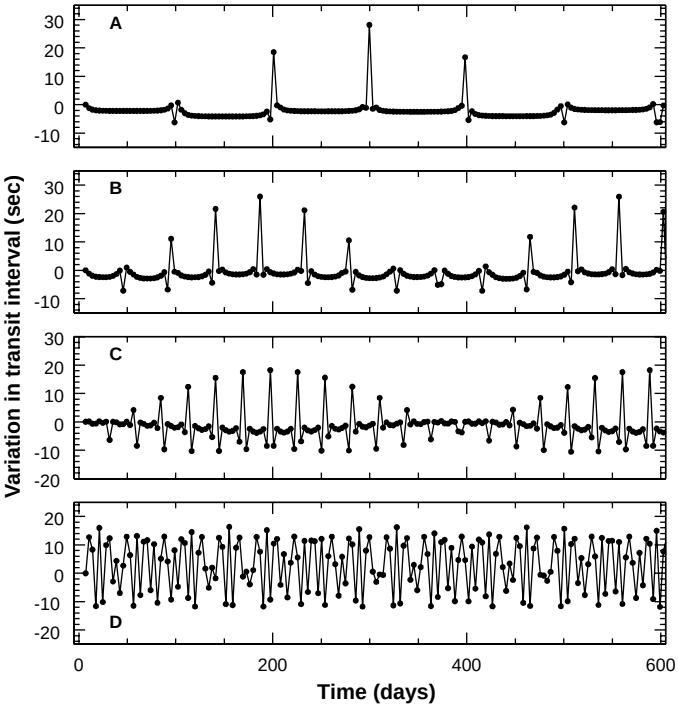


Fig. 2. Transit times of HD 209458b. The numerically determined variation of the interval between successive transit centers of HD 209458b as a function of time, with each panel showing the results for a different set of orbital parameters for a hypothetical second planet with a mass of $M_2 = 10^{-3}\ M_{\odot}$. In the simulations, HD 209458b has an initial orbital period $P_1 = 3.5248$ days and eccentricity $e_1 = 0.025$. The planets are assumed to be coplanar, with the system viewed edge-on. (**A**) The results for a perturbing planet with orbital period $P_2 = 99.8$ days and eccentricity $e_2 = 0.7$. (**B** to **D**) The results for perturbing planets with orbital periods P_2 of 46.4, 28.0, and 19.2 days and eccentricities e_2 of 0.5, 0.3, and 0.1, respectively. Radial-velocity measurements have ruled out the presence of such planets in the actual HD 209458b system (*20*).

parison of our different perturber simulations (Figs. 3 and 4) confirms that the magnitude of the effect is proportional to the perturber mass. For time spans shorter than the orbital period of the transiting planet, the magnitude of the effect is independent of the mass of the transiting planet (a result that follows from the equivalence of inertial and gravitational mass).

We estimate the variation in transit intervals for an inner transiting planet by integration of Lagrange's equations of planetary motion (*24*). Given a transiting planet with semimajor axis a_1 and period P_1, and a perturbing planet with semimajor axis a_2 (where $a_2 > a_1$), period P_2, and mass M_2, we find

$$\Delta t \simeq \frac{45\pi}{16}\left(\frac{M_2}{M_*}\right)P_1\alpha_e^3(1-\sqrt{2}\alpha_e^{3/2})^{-2} \quad (1)$$

where Δt is the magnitude of the typical variation of the interval between successive transit and $\alpha_e = \{(a_1)/[a_2(1-e_2)]\}$. Equation 1 was derived by assuming that the perturber follows a parabolic orbit with a periastron distance of $a_2(1-e_2)$. It underestimates the actual variation in transit period for small period ratios and best matches the numerical results for $e_2 \geq 0.3$. The transit-interval variations increase with P_1 when the period ratio of the two planets is held fixed. Thus, for a fixed timing accuracy, the detection of companions is easier for systems where the transiting planet is farther from the star. For a given perturbing planet mass, the timing variations are also larger for planets orbiting less massive stars.

The eccentricity of the outer planet can be estimated from the relative magnitudes of the variations $\Delta t_{max}/\Delta t_{min}$. If the period of the perturbing planet is much greater than that of the transiting planet, the final factor of Eq. 1 can be ignored. The resulting equation can be rearranged to provide an estimate of the mass of the perturbing planet

$$M_2 \simeq \frac{16}{45\pi}M_*\frac{\Delta t_{max}}{P_1}\left(\frac{P_2}{P_1}\right)^2(1-e_2)^3 \quad (2)$$

given an estimate of e_2.

Large excursions in transit-interval variation (Figs. 3 and 4) occur near small integer ratios of the orbital periods of the two planets, the locations of mean-motion resonances between the planets. In and near mean-motion resonances, the planets undergo larger oscillations of semimajor axis and eccentricity (*24*). The width in semimajor axis of each resonance region is proportional to $a_2(M_*/M_2)^{1/2}$ and grows rapidly with increasing eccentricity (*24*). The ranges of perturber periods (Figs. 3 and 4) in which the transit-interval variations are irregular correspond to dynamical chaos that results from the overlap of adjacent mean-motion resonances (*25*). We have excluded the ranges of perturber period in which this chaos results in short-term dynamical instability because such planetary configurations are unlikely to be found.

The two planets of the GJ 876 system provide an excellent example of a 2:1 mean-motion resonance in an extrasolar planetary system. Because of the resonant gravitational interaction between the two planets, the orbital periods of the inner and outer planet vary from 30.1 days to 31.1 days and from 60.0 days to 59.1 days, respectively, over a libration period of 550 days (*26*). Although careful photometric monitoring has excluded the possibility of transits of the inner planet of the GJ 876 system (*27*), such transit-interval variations would be seen if the inner planet transited.

In some cases, the libration period of a resonant system can be much longer than the interval of observation. Thus, the transit-interval variations due to the resonance cannot be easily discerned, even if the amplitude of the variation is large. GJ 876's short libration period, P_{lib}, results from the combination of its low-mass star ($M_* = 0.4\ M_\odot$) and massive planets ($M_1 = 2\ M_{jup}$, $M_2 = 4\ M_{jup}$), because $P_{lib} \propto P_1(M_*/M_2)^{1/2}$ (*26*, *24*) (and depends on the order of the resonance). The 10^4 day integrations used for Figs. 3 and 4 are long enough to sample a full libration period for the low-order resonances. A corresponding system with a solar-mass star, a 1-year orbital period for the inner planet, and Jupiter-mass planets in the 2:1 resonance would have a libration period of ~50 years. For Earth-mass planets, the libration period would be nearly 1000 years. These long-term effects would not be observable. However, the smaller transit-interval variations that occur on the time scale of the orbital periods of the two planets could be observed.

The feasibility of using this technique to detect additional planets in transiting systems

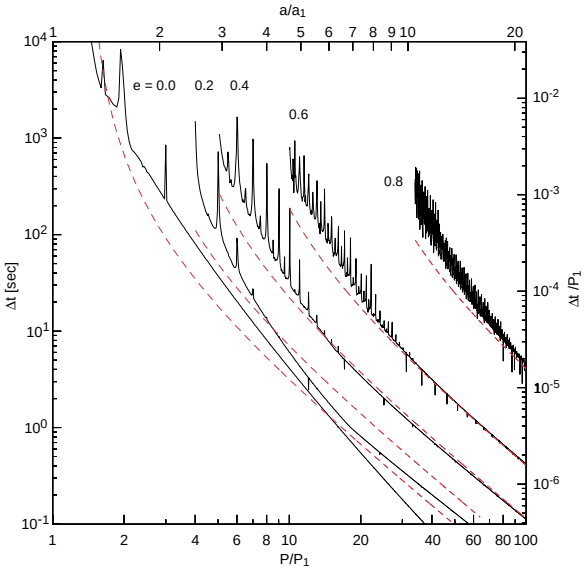


Fig. 3. Variations in the interval between successive transits of a planet with orbital period $P_1 = 3$ days, eccentricity $e_1 = 0.01$, and mass $M_1 = 10^{-3}\ M_\odot$, induced by a second planet with mass $M_2 = 10^{-3}\ M_\odot$. The planets are assumed to be coplanar. The black lines show the maximum transit-time variation as a function of orbital period (or semimajor axis) and eccentricity of the outer planet. The star has mass $M_* = M_\odot$. The red dashed lines show the estimate given by Eq. 1.

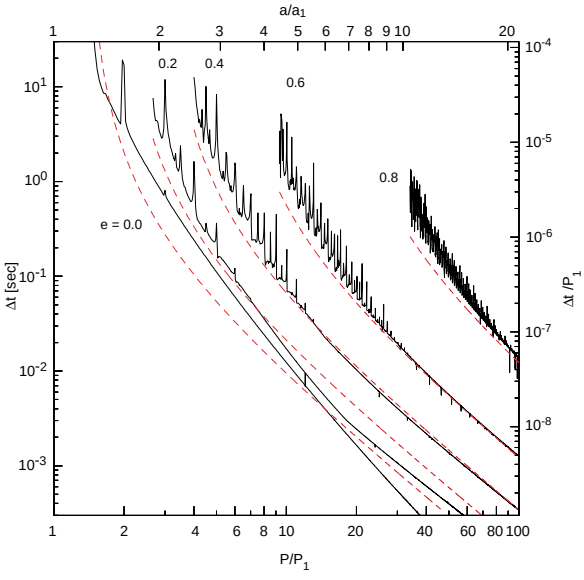


Fig. 4. Variations in the interval between successive transits of a planet with orbital period $P_1 = 3$ days, eccentricity $e_1 = 0.01$, and mass $M_1 = 10^{-3}\ M_\odot$, induced by a second planet with mass $M_2 = 3 \times 10^{-6}\ M_\odot$. The black lines show the maximum transit-time variation as a function of orbital period (or semimajor axis) and eccentricity of the outer planet. The star has mass $M_* = M_\odot$. The red dashed lines show the estimate given by Eq. 1.

depends on the physical and orbital properties of the multiple planet systems, the accuracy with which the times of transit can be measured, and the time span of the observations. Our expectations for systems with multiple Neptune-mass to Jupiter-mass planets are guided by recent discoveries, as well as by the giant planets in our solar system. Fourteen of 117 systems are known to have two or more planets. These have period ratios ranging from 2 to 150, and three are known or thought to be in mean-motion resonances. In the solar system, neighboring giant planets have period ratios of two to three. The eccentricities of the planets in the solar system are small, but in many of the extrasolar planetary systems the eccentricities are large, ranging up to $e = 0.9$.

Both the theory of terrestrial-planet formation and available observations suggest that the typical terrestrial-planet system will have a configuration that results in variations in the transit interval of hundreds of seconds over time spans of months to years. Because the escape velocity from the surface of a terrestrial planet is smaller than the escape velocity from a solar-mass star at ~1 AU, all the solid material in the region from a few tenths to one or two AU must either accrete into planets or fall onto the star. The resulting planets are as closely spaced as dynamical stability permits (*28*, *29*). As observed in the solar system and in the pulsar planet system, the period ratios are of order two.

Assuming that the observations during the ingress and egress of a transit are well sampled, the error σ_{tc} (due to photon statistics alone) associated with the measured time of the center of a transit of duration t_T is given by

$$\frac{\sigma t_c}{t_T} \sim (\Gamma t_T)^{-1/2} \rho^{-3/2} \qquad (3)$$

where Γ is the photon count rate of the star and $\rho = R_p/R_*$ is the ratio of the planet radius to the stellar radius. Kepler (*15*), a NASA Discovery mission, will monitor 100,000 A-M dwarf stars brighter than apparent magnitude $V = 14$, looking for transits of terrestrial-sized planets in the habitable zone (*30*). The smaller mission, COROT, will monitor 12,000 dwarf stars with apparent magnitudes of $V = 11$ to $V = 16$, also searching for terrestrial-mass planets (*31*). For Kepler, with its 0.95-m-diameter aperture, $\Gamma = 7.8 \times 10^8\ 10^{-0.4(V-12)}\ h^{-1}$, for a star of apparent magnitude V. For a Jupiter-sized planet in a 1-year orbit about a solar-mass star with $V = 12$, we find $\sigma_{tc} \sim 20$ s ($\rho \sim 0.1$, $t_T \sim 13$ h). For a terrestrial-sized planet, $\sigma_{tc} \sim 500$ s ($\rho \sim 0.01$). These accuracies suggest that the transit-period variations due to the gravitational influence on each other of Earth-mass planets with small period ratios can be detected by Kepler. The observing cadence for Kepler will be limited by the rate at which the data are transmitted to Earth. This will nominally be 15 min but will be reduced to 1 min for likely transit candidates. For brighter stars observed with large-aperture ground-based telescopes, the presence of more distant additional planets can be detected. For example, observing a $V = 9$ star with a 6.5-m-aperture telescope, $\sigma_{tc} \sim 0.2$ s. For such a star, the photon noise exceeds the errors due to atmospheric scintillation after ~30 s of accumulated integration (*32*). Also, microvariability at the level of 10 to 50 parts per million is expected from p-mode oscillations such as those seen in the sun, with characteristic periods of 300 s. This noise source is competitive with shot noise in 8-m-class telescopes for $V = 14$ stars.

One would like to measure the density of a transiting planet to determine whether it is a rocky terrestrial planet such as Earth, an ice/water giant such as Neptune, or a gas giant planet such as Jupiter. Unfortunately, when a transiting planet is perturbed by a nontransiting planet, photometry yields the radius of the transiting planet but not its mass, whereas the transit times yield the mass of the perturbing planet but not its radius. Thus, neither planet's density can be determined.

In cases in which two or more planets transit their star, the masses and radii of each planet can be estimated, allowing density determinations for all. For such systems, radial-velocity observations will not be necessary to determine their masses and densities. Many of the target stars in future surveys will be too faint to allow for radial-velocity detections of terrestrial-mass planets, even with technical improvements to the accuracy of such measurements. Thus, the transit-timing technique may be the only means to estimate the mass and density of terrestrial-mass planets. If one sees transits of one planet in a multiple-planet system, what is the probability of seeing transits of a second planet? Assuming that the orbits of the two planets are coplanar, the probability that the second planet also transits is a_1/a_2 if $a_2 > a_1$ (if $a_2 < a_1$, transits of the second planet are assured in the case of coplanar orbits). If the orbits of the two planets are mutually inclined and one planet transits the center of the star, the probability that the second planet also transits is

$$P_{t_2} = \frac{2}{\pi} \arcsin\left(\frac{R_*}{a_2 \sin i'}\right) \qquad (4)$$

where i' is the mutual inclination between the orbits of the two planets rather than the sky-plane inclination. This assumes $\sin i' > R_*/a_2$; otherwise, transits of the second planet are certain. For a solar-radius star, a mutual inclination of a few degrees, and a semimajor axis $a_2 \sim 1$ AU, the probability is roughly 10%. Thus, such double-transiting systems are likely to be found.

The Kepler team plans to search for transits with a consistent period, depth, and duration (*15*). We caution that any detection algorithm that looks for evidence of periodic transits in the photometry must allow for variations in the transit period due to the perturbations from unseen planets. An overly restrictive test for periodicity might reject some of the most interesting and informative planetary systems. The GJ 876 system discussed above, an extreme case, might be rejected as only quasiperiodic.

We also caution that experiments looking for gradual transit-period variations due to slow secular trends must allow for the possibility of substantial short-term variations in the transit period. Transit-period variations due to orbital precession (*16*) or to tidal interaction with a star (*17*) require many years to elapse before they can be detected; without careful monitoring in the interim to determine the short-term variations, the results of such experiments could easily be misinterpreted.

References and Notes

1. J. Schneider (2004); www.obspm.fr/encycl/catalog.html.
2. I. A. Bond *et al.*, *Astrophys. J.* **606**, L155 (2004).
3. A. Wolszczan, D. A. Frail, *Nature* **355**, 145 (1992).
4. D. C. Backer, R. S. Foster, S. Sallmen, *Nature* **365**, 817 (1993).
5. M. Mayor, D. Queloz, *Nature* **378**, 355 (1995).
6. G. W. Marcy, R. P. Butler, *Astrophys. J.* **464**, L147 (1996).
7. D. Charbonneau, T. M. Brown, D. W. Latham, M. Mayor, *Astrophys. J.* **529**, L45 (2000).
8. G. W. Henry, G. W. Marcy, R. P. Butler, S. S. Vogt, *Astrophys. J.* **529**, L41 (2000).
9. A. Udalski *et al.*, *Acta Astron.* **52**, 1 (2002).
10. A. Udalski *et al.*, *Acta Astron.* **52**, 115 (2002).
11. A. Udalski *et al.*, *Acta Astron.* **53**, 133 (2003).
12. B. E. McArthur *et al.*, *Astrophys. J.* **614**, L81 (2004).
13. P. Butler *et al.*, *ArXiv Astrophysics e-prints*, astro-ph/0408587 (2004).
14. R. Narayan, A. Cumming, D. N. C. Lin, *ArXiv Astrophysics e-prints*, astro-ph/0409766 (2004).
15. W. J. Borucki *et al.*, in *Future EUV/UV and Visible Space Astrophysics Missions and Instrumentation*, J. C. Blades, O. H. W. Siegmund, Eds. Proceedings of the SPIE, Vol. 4854, pp. 129–140 (2003).
16. J. Miralda-Escudé, *Astrophys. J.* **564**, 1019 (2002).
17. D. D. Sasselov, *Astrophys. J.* **596**, 1327 (2003).
18. E. W. Brown, *Mon. Not. R. Astron. Soc.* **97**, 62 (1936).
19. S. Soderhjelm, *Astron. Astroph.* **42**, 229 (1975).
20. P. Bodenheimer, G. Laughlin, D. N. C. Lin, *Astrophys. J.* **592**, 555 (2003).
21. E. Agol, J. Steffen, R. Sari, W. Clarkson, *ArXiv Astrophysics e-prints*, astro-ph/0412032 (2004).
22. J. Stoer, R. Bulirsch, *Introduction to Numerical Analysis* (Springer-Verlag, New York, 1980).
23. T. R. Quinn, S. Tremaine, M. Duncan, *Astron. J.* **101**, 2287 (1991).
24. C. D. Murray, S. F. Dermott, *Solar System Dynamics* (Cambridge Univ. Press, Cambridge, 1999).
25. J. Wisdom, *Astron. J.* **85**, 1122 (1980).
26. G. Laughlin, J. E. Chambers, *Astrophys. J.* **551**, L109 (2001).
27. G. Laughlin (2004); www.transitsearch.org.
28. J. E. Chambers, G. W. Wetherill, *Icarus* **136**, 304 (1998).
29. E. Kokubo, S. Ida, *Icarus* **131**, 171 (1998).
30. J. F. Kasting, D. P. Whitmire, R. T. Reynolds, *Icarus* **101**, 108 (1993).
31. C. Catala *et al.*, *Astronomical Society of the Pacific Conference Series* **76**, 426 (1995).
32. R. L. Gilliland *et al.*, *Astron. J.* **106**, 2441 (1993).

33. This work was supported in part by NSF grant PHY99-07949 and by NASA grant NAG5-9678. This research was supported by the Natural Sciences and Engineering Research Council of Canada and by the Canada Research Chair program. N.W.M. is a Canada Research Chair. We are grateful to the Kavli Institute for Theoretical Physics, where much of this investigation was carried out. We thank S. Gaudi and J. Winn for helpful discussions and careful reviews of the manuscript.

23 November 2004; accepted 21 January 2005
10.1126/science.1107822

Unveiling Extensive Clouds of Dark Gas in the Solar Neighborhood

Isabelle A. Grenier,[1]* Jean-Marc Casandjian,[1,2] Régis Terrier[3]

From the comparison of interstellar gas tracers in the solar neighborhood (HI and CO lines from the atomic and molecular gas, dust thermal emission, and γ rays from cosmic-ray interactions with gas), we unveil vast clouds of cold dust and dark gas, invisible in HI and CO but detected in γ rays. They surround all the nearby CO clouds and bridge the dense cores to broader atomic clouds, thus providing a key link in the evolution of interstellar clouds. The relation between the masses in the molecular, dark, and atomic phases in the local clouds implies a dark gas mass in the Milky Way comparable to the molecular one.

The census of atomic and molecular gas in the Milky Way mainly comes from the surveys of the 21-cm-HI and 2.6-mm-CO lines. H_2 molecules are not directly observed in the cold molecular clouds, but the velocity-integrated CO intensity, W(CO), empirically scales to the N(H_2) column density by the X = N(H_2)/W(CO) factor. Cold dust, near 16 to 20 K, is mixed in both gas phases, and its thermal emission dominates near 3000 GHz. Colder dust, around 10 K, is seen down to 94 GHz. Dust column densities are traced by optical extinction and reddening measurements. γ rays with energies above 100 MeV are produced by cosmic rays interacting with interstellar gas and radiation, so they can effectively trace the gas mass independent of its chemical or thermodynamic state throughout the Galaxy. These tracers have different limitations and comparing them is essential to map and weigh the main gas reserves of the Milky Way. We have compared them over the whole sky at $5° \leq |b| \leq 80°$, where local clouds dominate and can be mapped in detail.

The HI column-density map, N(HI), combines the 0.5°-resolution Leiden/Dwingeloo survey (*1*) for declinations >–30° with the 1°-resolution data farther south (*2*). The HI lines were assumed to be optically thin. The W(CO) map comes from the Center for Astrophysics compilation at $|b| \leq 32°$ (*3*). The regions outside the survey should be free of bright CO emission (*3*, *4*). The dust intensity map at 3000 GHz, I_{3000}, comes from InfraRed Astronomical Satellite (IRAS) maps calibrated with Diffuse InfraRed Background Experiment (DIRBE) data, with the zodiacal light and point sources removed (*5*). Correcting I_{3000} for color temperature variations yields dust column densities. The reddening [E(B-V)] map (*5*), in magnitudes (mag.), comes from I_{3000} corrected to a temperature of 18.2 K with the use of DIRBE 100- to 240-μm ratios and calibrated to the reddening of elliptical galaxies with a 16% precision. It is consistent with star estimates from the Sloan Digital Sky Survey for E(B-V) < 0.5 (*6*). Only 6% of the region under study exhibits larger reddening. The dust intensity map at 94 GHz, I_{94}, was constructed (*7*) with two modified Planck $\nu^{1.67}$ B_ν (9.4 K) and $\nu^{2.70}$ B_ν (16.2 K) emissivities to match the Cosmic Background Explorer (COBE) Differential Microwave Radiometers (DMR) and Far InfraRed Absolute Spectrophotometer (FIRAS) data. It was successfully compared with the Wilkinson Microwave Anisotropy Probe (WMAP) data at 94 GHz (*8*).

In γ rays, the Energetic Gamma Ray Experiment Telescope (EGRET) archival database provides a 9-year photon count map and an exposure map, ε, for events above 100 MeV and within 30° of the telescope axis. The inverse-Compton (IC) intensity map, I_{IC}, integrated above 100 MeV, comes from the GALPROP 45-600202 model for cosmic-ray propagation in the Galaxy (*9*), and its estimate of the radiation field, which includes the cosmological microwave background (CMB), stellar radiation, and dust radiation with peaks near 100 and 10 μm. Given the EGRET exposure and point-spread function (PSF) above 100 MeV, we produced a map, N_{sou} (SOU), of the photon count contribution from the 138 persistent point sources of the third EGRET catalog (*10*) with a detection significance >5σ at $|b| < 30°$ and >4σ at higher latitudes. We used their 4-year average fluxes. Flaring sources not detected in the 4-year data were not included. The mid-latitude threshold was raised from 4σ to 5σ because the addition of substantial interstellar emission may cast doubts on the importance of the faint sources spatially correlated with gas tracers (*11*). We discarded 24 such sources.

Because energetic cosmic rays can penetrate all gas phases and the γ rays radiate easily away, the γ-ray intensity can be modeled as a linear combination of gas tracers, the IC map, the extragalactic background intensity (I_E), and the point sources. The contribution from the ionized hydrogen is neglected. The modeled photon counts in 0.5° × 0.5° bins, $N_p(l,b) = [q_{HI}\,\mathrm{N(HI)}(l,b) + q_{CO}\,\mathrm{W(CO)}(l,b) + q_{dust}\,I_{dust}(l,b) + q_{IC}\,\mathrm{IIC}(l,b) + I_E]\,\varepsilon(l,b) + q_{SOU}\,N_{sou}(l,b)$, are convolved with the PSF for an input $E^{-2.1}$ spectrum. I_{dust} refers to any of the dust maps. The emissivities, q, were fitted to the γ-ray data for different sets of tracers by means of a maximum-likelihood test with Poisson statistics (*12*) (Table 1). Consistent results were obtained at $|b| \geq 5°$ and 10°, indicating that the derived emissivities are local-medium averages. We checked that they are not biased by the lack of cosmic-ray gradient in the Galactic disk. Different point-source data sets yield <10% variations in q_{dust} and <1% variations in the other parameters. We used complete and residual dust maps; the latter were obtained by removing the part linearly correlated with N(HI) and W(CO) from a χ^2 fit using equal area pixels (*13*). Both give the same maximum-likelihood values and q_{dust}. The residual maps ($I_{3000res}$, $E(B\text{-}V)_{res}$, and I_{94res}) were used to evaluate "dust-free" q_{HI} and q_{CO} γ-ray emissivities.

Compared with earlier studies that used only N(HI) (*5*, *14*), both the I_{94res} and $E(B\text{-}V)_{res}$ maps confirm the existence of extensive coherent regions with apparently large dust-to-gas excesses. Additional gas, seen in γ rays but not in radio, complements the dust. The excesses form large halos systematically surrounding all nearby CO clouds (Fig. 1). By filtering out the large-scale HI contribution from the Galactic disk, we have isolated nearby HI clouds that are closely associated with the excess clouds (Fig. 2). The latter bridge the CO cores to the diffuse HI clouds. They are much more prominent at low frequency than at 3000 GHz, confirming the cold temperatures indicated by the DIRBE, FIRAS, and WMAP data, and they largely contribute to the "anomalous microwave emission" (*5*, *7*, *8*, *14*, *15*). They do not spatially correlate with free-free emission.

Even though dust is mixed in the atomic and molecular phases, the maximum-likelihood tests to the γ-ray data reject the I_{3000} dust map as a single tracer of the whole gas. The N(HI)+W(CO) model yields a much better fit [$2\ln(\lambda) > 13{,}530$]. The maximum-likelihood increase from the N(HI)+W(CO)

[1]Astrophysique Interactions Multi-échelles (CEA/Universite Paris 7/CNRS), Commissanat á l'Energie Atomique, DSM/DAPNIA, Service d'Astrophysique, CEA, Saclay, 91191 Gif sur Yvette, France. [2]Grand Accélérateur National d'Ions Lourds (CEA/CNRS), boulevard Henri Becquerel, BP 55027, 14076 Caen, France. [3]Astroparticules et Cosmologie (CNRS/Université Paris 7/CEA), 11 place Marcellin Berthelot, 75005 Paris, France.

*To whom correspondence should be addressed. E-mail: isabelle.grenier@cea.fr

to the N(HI)+W(CO)+I_{3000} model mostly comes from the Cygnus, Ophiuchus, and Orion A and B clouds. Little γ-ray emission is associated with the I_{3000} map on larger scales. The γ-ray fit largely improves (Fig. 3 and fig. S1) when using the colder dust tracer I_{94} and the dust column densities as traced by the E(B-V) reddening [see the 2ln(λ) values in Table 1].

The q_{HI} emissivity found at $5° \le |b| \le 80°$ is consistent with the solar circle average (±SD) of $(1.60 \pm 0.02) \times 10^{-26}$ s^{-1} sr^{-1} (*16*) and 15% lower than those obtained near major local clouds when neglecting the IC emission (*17*). The extragalactic intensity agrees with that measured with the EGRET interstellar model: $(14.5 \pm 0.5) \times 10^{-6}$ cm^{-2} s^{-1} sr^{-1} (*18*, *19*). It is larger than the $(11.10 \pm 0.12 \pm 2.2) \times 10^{-6}$ cm^{-2} s^{-1} sr^{-1} GALPROP estimate (*20*) because of the 30% lower IC intensity than that in GALPROP (*9*). Provided cosmic rays uniformly pervade the HI and H_2 phases, the N(HI)+W(CO) model yields an average $X_\gamma = q_{CO}/(2q_{HI}) = (1.74 \pm 0.03) \times 10^{20}$ molecules (mol.) cm^{-2} K^{-1} km^{-1} s in the solar neighborhood, fully consistent with the factor $X = (1.8 \pm 0.3) \times 10^{20}$ mol. cm^{-2} K^{-1} km^{-1} s obtained with the temperature-corrected I'_{3000} map and a smoothed I'_{3000}/N(HI) ratio to measure the total gas (*3*). It is close to the Galactic averages of $(1.9 \pm 0.2) \times 10^{20}$ mol. cm^{-2} K^{-1} km^{-1} s (*16*) and $(1.56 \pm 0.05) \times 10^{20}$ mol. cm^{-2} K^{-1} km^{-1} s (*19*) derived from γ rays and different gas models.

The considerable improvement in the γ-ray fits when adding the I_{94} or E(B-V) dust tracers demonstrates that γ rays are produced in the cold dust halos, in excess of those born in the HI and CO gas or in the warmer dust in these directions. Splitting the sky in three longitude intervals did not change the result, so one particular region does not drive the fit. This bright emission, comparable to the entire I_{IC} map, is not due to IC emission from the colder dust in the cloud halos. It contributes only 0.6 and 0.9% of the CMB and warmer dust emissions [for a local electron spectrum $I_e = 708$ (E/1 GeV)$^{-3.3}$ m^{-2} s^{-1} sr^{-1} GeV^{-1} (*21*) and two dust components with $\nu^{1.67}$ B_ν (9.4 K) and $\nu^{2.70}$ B_ν (16.2 K) emissivities scaled to match typical fluxes of 1.5 MJy/sr at 500 GHz (FIRAS) and 15 MJy/sr at 3000 GHz (IRAS)]. The lack of correlation with the 408 MHz synchrotron map shows that GeV cosmic rays are not efficiently trapped in these cloud halos. Their distribution couples to gas density (*19*) and spiral arms (*17*), but their diffusion properties imply coupling lengths much larger than the cloud sizes: 1.76 ± 0.20 kpc for nuclei (*19*) and several hundred parsecs for the highest energy electrons with large radiative losses (*22*). Measurements of q_{HI} indeed yield equivalent cosmic-ray densities within 1 kpc from the Sun (*17*). Because the local dust mass is only 1% of the gas one, hadronic interactions with grains are negligible. The γ rays therefore trace the dark gas forming these cloud halos, a gas invisible in HI, CO, and free-free emission.

Assuming the same cosmic-ray flux in the diffuse dark and HI gas, the γ-ray intensities give dark-gas column densities, NH_{dark} (Fig. 4 and fig. S2). The E(B-V) and I_{94} tracers, total and residual maps, give consistent NH_{dark} column densities within 30% over most regions. Fitting cloud complexes individually may lead

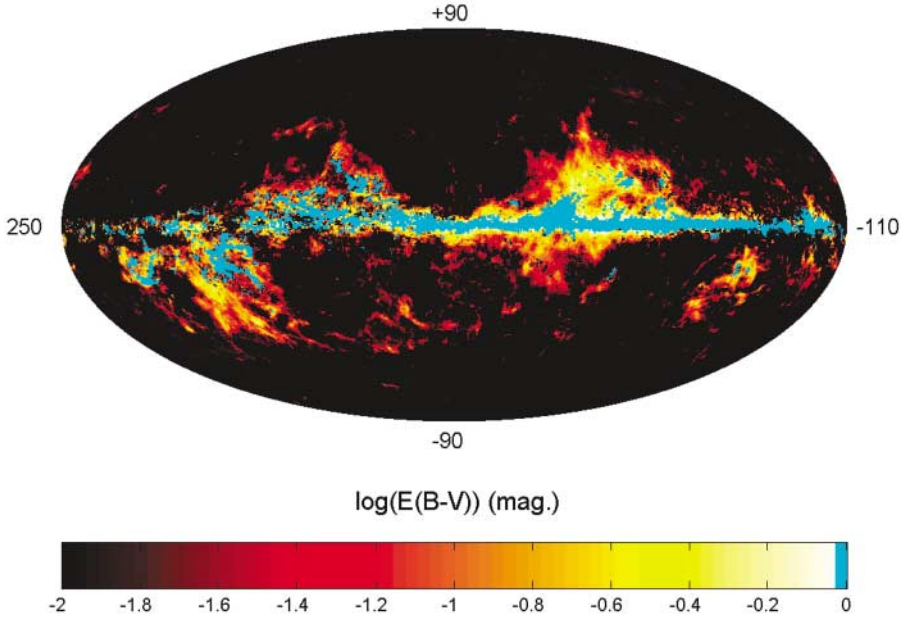


Fig. 1. Map, in Galactic coordinates centered on $l = 70°$, of the excess dust reddening found above that linearly correlated with the integrated HI and CO line intensities. The 94-GHz emission map shows the same excesses. CO intensities above 4 K km s^{-1} are overlaid in cyan. The dust excesses form extended halos around all CO clouds, the bright ones as well as the fainter CO cloudlets that are not overlaid. This dust spreads in a dark gas, not seen in HI and CO but detected in γ rays.

Table 1. Best fits to the γ-ray data at $5° \le |b| \le 80°$ and $10° \le |b| \le 80°$, for the complete and residual (in bold and italics) dust maps. The log-likelihood ratios 2.ln(λ), measuring the fit improvement, are given with respect to the N(HI)+W(CO) gas model and are distributed as a χ^2 with one degree of freedom (*12*). The errors are statistical, not including systematic uncertainties [10% in the EGRET exposure (*10*), 20% in the true N(HI) (*32*) and dust maps (*5*), and 20% in the "dust-free" q_{CO} because of cloud-to-cloud variations in the dust residuals toward the CO clouds].

2.ln(λ)	q_{HI} (10^{-26} s^{-1} sr^{-1})	q_{CO} (10^{-6} cm^{-2} sr^{-1} K^{-1} km^{-1})	q_{EBV} (10^{-6} mag^{-1} cm^{-2} s^{-1} sr^{-1})	q_{I94} (10^{-4} mK^{-1} cm^{-2} s^{-1} sr^{-1})	q_{I3000} (10^{-7} MJy^{-1} cm^{-2} s^{-1})	q_{IC} (10^{-1})	q_{SOU}	I_E (10^{-6} cm^{-2} s^{-1} sr^{-1})
				$5° \le \|b\| \le 80°$				
0	1.62 ± 0.01	5.65 ± 0.10				7.46 ± 0.06	1.01 ± 0.01	13.21 ± 0.09
241.8	***1.63 ± 0.01***	***5.54 ± 0.09***			***2.83 ± 0.20***	***6.54 ± 0.09***	***1.00 ± 0.01***	***13.93 ± 0.10***
1298.0	***1.64 ± 0.01***	***5.82 ± 0.09***		***4.27 ± 0.12***		***6.39 ± 0.07***	***0.96 ± 0.01***	***13.96 ± 0.09***
1556.7	***1.65 ± 0.01***	***5.91 ± 0.09***	***57.5 ± 1.5***			***6.43 ± 0.07***	***0.96 ± 0.01***	***13.91 ± 0.09***
241.8	1.49 ± 0.02	4.84 ± 0.11			2.83 ± 0.21	6.54 ± 0.09	1.00 ± 0.01	13.74 ± 0.10
1298.0	0.77 ± 0.02	1.63 ± 0.15		4.27 ± 0.10		6.39 ± 0.07	0.96 ± 0.01	14.30 ± 0.09
1556.7	0.71 ± 0.04	0.53 ± 0.54	57.5 ± 3.7			6.43 ± 0.07	0.96 ± 0.01	14.45 ± 0.09
				$10° \le \|b\| \le 80°$				
0	1.74 ± 0.02	5.31 ± 0.12				7.90 ± 0.09	0.98 ± 0.01	12.40 ± 0.10
284.4	***1.74 ± 0.02***	***5.28 ± 0.12***			***4.04 ± 0.25***	***7.07 ± 0.10***	***0.97 ± 0.01***	***13.03 ± 0.10***
957.9	***1.75 ± 0.02***	***5.70 ± 0.12***		***4.32 ± 0.14***		***7.25 ± 0.09***	***0.91 ± 0.01***	***12.84 ± 0.10***
1048.6	***1.76 ± 0.02***	***5.60 ± 0.12***	***54.4 ± 1.7***			***7.17 ± 0.09***	***0.91 ± 0.01***	***12.83 ± 0.10***

to 20% variations, but it first requires a careful reanalysis of the point-source fluxes. The gas-to-dust ratio deduced from the amount of dark gas is consistent with the average ratio found in the local CO clouds in (*13*): $[NH/E(B\text{-}V)]_{dark} = (3.5 \pm 0.2) \times 10^{21}$ cm^{-2} mag^{-1} and $[NH/E(B\text{-}V)]_{CO} = (3.72 \pm 0.28) \times 10^{21}$ cm^{-2} mag^{-1}. Both are lower than the average of $(6.17 \pm 0.22) \times 10^{21}$ cm^{-2} mag^{-1} found in the local HI gas, close to the canonical 5.8×10^{21} cm^{-2} mag^{-1} value (*23*), but reflect a metallicity close to normal.

We calculated the masses in the molecular, dark, and atomic phases in seven regions from 140 to 450 pc from the Sun (Table 2), with the use of X_γ, a mean weight per atom of 1.4 times the hydrogen mass, and the CO and HI data in Figs. 1 and 2. We added the masses of CO clouds in the MBM 53-54-55 and Chamaeleon complexes (*24*, *25*) that are not covered by our data (*4*). Because the three phases are spatially connected, their relative masses are independent of distance. The dark-gas mass increases similarly with the independent H_2 and HI mass estimates, approximately as $M^{0.4}$ and $M^{0.5}$, so its relative contribution decreases from small to massive clouds (Table 2 and fig. S3). The sample is small but spans the mass range most typical of Galactic clouds, from translucent 10^3–solar mass ($M_\odot$) clouds to giant complexes $>10^5$ $M_\odot$. In smaller cirrus clouds, there are also comparable masses of HI and dark gas (inferred from dust 3000 GHz emission) without CO detection (*26*). For a uniform X factor, the mass spectrum of Galactic CO clouds follows a power-law $dN/dM_{H2} \propto M_{H2}^{-\alpha}$, where $\alpha = 1.6$ to 1.9 [$\alpha = 1.67 \pm 0.25$ in the first quadrant in the range of $10^{4.5}$ to $10^{6.5}$ $M_\odot$ (*27*) and $\alpha = 1.80 \pm 0.03$ from 600 to 10^6 $M_\odot$ in the outer Galaxy (*28*)]. The best χ^2 power-law fit of M_{dark} with M_{H2} in the local regions, which matches the lowest dark masses and underestimates those in both Aquila regions by a factor of 4, is $M_{dark} = (457 \pm 68) \times (M_{H2}^{0.36 \pm 0.03})$ in solar masses (fig. S3). Together with the CO mass spectrum, it implies a minimum dark-to-molecular mass fraction in the Galaxy of 50 to 100% in the interval of 10^3 to 10^6 $M_\odot$, for $\alpha = 1.6$ to 1.9, respectively. Locally, $M_{dark}/M_{H2} \approx 30\% \approx M_{dark}/M_{HI}$, which should be taken as lower limits because the local sample is strongly biased to high masses with less dark gas (*27*). So the dark-cloud mass in the Galaxy should compare with the molecular one. It strengthens the possibility that cold H_2 substantially contribute to the dark baryonic matter (*29*).

Whether the dark gas is gravitationally bound and dominated by cold HI or H_2 is unclear. H_2 number fractions of 49 to 76% have been found in ultraviolet absorption lines toward three stars in the Chamaeleon clouds, with an H_2 formation time scale of 20 million years less than the photodissociation one (*30*).

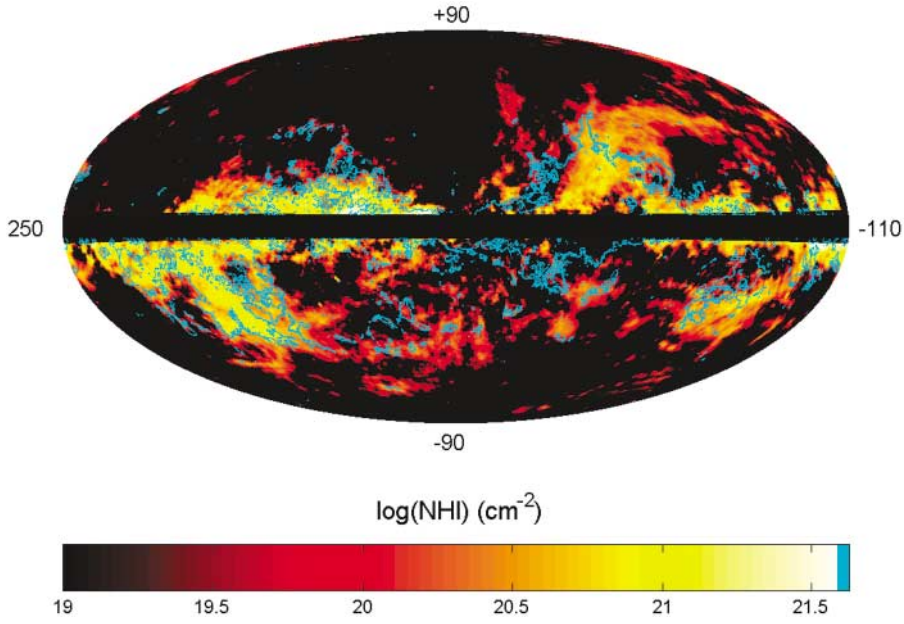


Fig. 2. Map, in Galactic coordinates centered on $l = 70°$, of the HI column densities in nearby atomic clouds isolated by subtracting the Galactic disk emission from the best-fit cosecant law about the warped plane. The outer contour of the dust excesses of Fig. 1 is overlaid in cyan, showing a remarkable association with the atomic clouds.

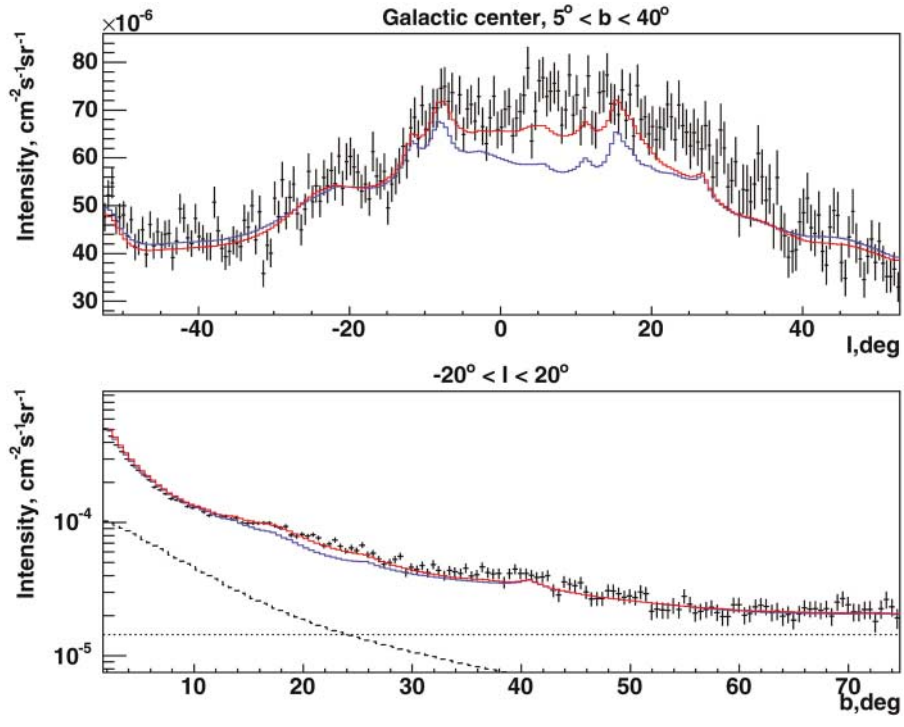


Fig. 3. Longitude (**top**) and latitude (**bottom**) profiles of the observed γ-ray intensity in the Aquila-Ophiuchus-Libra region versus the N(HI)+W(CO) gas model (blue) and the N(HI)+W(CO)+E(B-V) model (red). The dashed and dotted curves (bottom) outline the IC and extragalactic background intensities, respectively. Error bars show mean ± SD.

Larger scales are probed in CO (0.3 pc) and γ rays (1.4 pc). In the direction with no CO detection, the dark gas could contribute half of the H_2 seen in absorption (table S1). The low $[NH/E(B\text{-}V)]_{dark}$ ratio, close to the CO one, also supports a high H_2 abundance. There could be a sharp threshold in extinction ($A_v > 0.25$ mag) before CO is detectable (*31*), but the 0.2- to 1.5-magnitude extinction in the dark clouds provides enough shielding for the molecules. Assuming pure H_2 halos, we find potential CO intensities of 0.5 to 1.5 K km s^{-1}

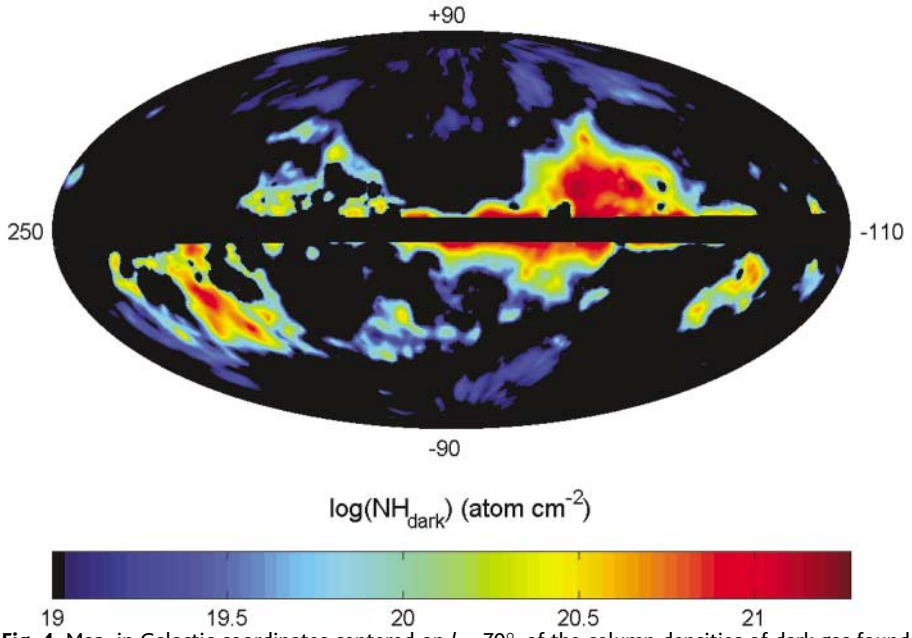


Fig. 4. Map, in Galactic coordinates centered on $l = 70°$, of the column densities of dark gas found in the dust halos, as measured from their γ-ray intensity with the reddening map. This gas complements that visible in HI and CO. The two dust tracers [E(B-V) and 94-GHz emission] yield consistent values within 30% over most regions.

Table 2. Total mass and H_2:dark:HI mass proportions in the CO, dark, and atomic phases in nearby regions.

Region	Longitudes, latitudes	Distance (pc)	Total mass (10^5 $M_\odot$)	H_2:dark:HI
Cepheus-Cassiopeia-Polaris	[80,165] [5, 50]	300	5.4	7.5:1:7.3
Orion	[–163, –134] [–40, –5]	450	4.8	8.6:1:7.3
Aquila-Ophiuchus-Libra	[–70, 50] [5, 60]	140	1.6	0.8:1:1.0
Taurus-Perseus-Triangulum	[140, 197] [–60, –5]	140	1.4	2.3:1:3.1
Aquila-Sagittarius	[–36, 50] [–50, –5]	140	0.43	0.2:1:0.5
Chamaeleon	[–90, –35] [–50, –8]	160	0.40	0.9:1:3.7
Pegasus	[50, 140] [–60, –26]	150	0.16	0.7:1:3.4

in most places, close to the sensitivity of the large-scale surveys for small-size cloudlets, and higher values in the Aquila-Ophiuchus-Libra (2 to 4 K km s^{-1}) and Taurus-Perseus-Triangulum (2 K km s^{-1}) regions where only faint scattered CO cloudlets were detected. On the other hand, the dark clouds may be remnants of the atomic clouds that condensed to form the CO cores or they may be rich in photodissociated gas torn away in the ambient turmoil. Cold HI, not properly accounted for in emission because of its mixing with warm HI lines, has been observed with and without CO. It shows abundant structure in self-absorption, but its extent at medium latitude is unknown. NHI_{warm} and NHI_{cold} column densities have been measured against 65 radio sources (*32*), 22 of which are seen through dark clouds. A weak correlation was found between $NH_{halo}/\sin|b|$ and $NHI_{cold}/\sin|b|$, with a correlation coefficient of 0.67 ± 0.09. None was found with NHI_{warm}. So, whereas the dark gas provides a key link in the evolution between the atomic and molecular clouds, the nature of its major constituent is unknown.

The addition of substantial mass to the interstellar γ-ray emission implies that the cosmic-ray densities estimates are currently overestimated in regions rich in dark gas, away from the giant molecular complexes. This should steepen the cosmic-ray gradient with Galactocentric radius and enhance the arm-interarm contrast in cosmic rays. Yet, resolving the dark clouds in the Galactic disk, with dust temperature gradients near bright OB associations along crowded lines of sight and poor angular resolution in γ rays, is a difficult but important challenge. The presence of dark clouds to high latitudes also implies significant revisions of the EGRET catalog of γ-ray sources.

References and Notes

1. D. Hartmann, W. B. Burton, *Atlas of Galactic Neutral Hydrogen* (Cambridge Univ. Press, Cambridge, 1997).
2. J. M. Dickey, F. J. Lockman, *Annu. Rev. Astron. Astrophys.* **28**, 215 (1990).
3. T. M. Dame, D. Hartmann, P. Thaddeus, *Astrophys. J.* **547**, 792 (2001).
4. T. M. Dame, P. Thaddeus, in *Milky Way Surveys: The Structure and Evolution of Our Galaxy*, D. Clemens, T. Brainerd, Eds. (ASP Conference Series, Astronomical Society of the Pacific, San Francisco, CA, 2004), pp. 66–72.
5. D. J. Schlegel, D. P. Finkbeiner, M. Davis, *Astrophys. J.* **500**, 525 (1998).
6. D. P. Finkbeiner *et al.*, *Bull. Am. Astron. Soc.* **36**, 2 (2004).
7. D. P. Finkbeiner, M. Davis, D. J. Schlegel, *Astrophys. J.* **524**, 867 (1999).
8. D. P. Finkbeiner, *Astrophys. J.* **614**, 186 (2004).
9. A. W. Strong, I. V. Moskalenko, O. Reimer, *Astrophys. J.* **613**, 962 (2004).
10. R. C. Hartman *et al.*, *Astrophys. J. Suppl. Ser.* **123**, 79 (1999).
11. I. Grenier, *Astron. Astrophys.* **364**, L93 (2000).
12. W. T. Eadie, D. Drijard, F. James, M. Roos, B. Sadoulet, *Statistical Methods in Experimental Physics* (North-Holland, Amsterdam, 1971).
13. The best linear fits to the dust data at $5° \leq |b| \leq 80°$, in equal area pixels of 1.6×10^{-5} sr, follow. We used $\sigma_{HI} = 3 \times 10^{18}$ cm^{-2}, $\sigma_{CO} = 0.6$ K km s^{-1}, and $\sigma_{dust}/I_{dust} = 0.2$. N(HI) is in units of centimeters^{-2}, W(CO) is in kelvin kilometer seconds, E(B-V) is in magnitude, I_{94} is in millikelvins, and I_{3000} in megajansky per steradian. The following equations show the best linear fits:

$$E(B\text{-}V) = (1.62 \pm 0.06) \times 10^{-22}\ N(HI) + (9.4 \pm 0.7) \times 10^{-2}\ W(CO) - (9.4 \pm 0.3) \times 10^{-3}$$

$$I_{94} = (2.04 \pm 0.04) \times 10^{-23}\ N(HI) + (9.8 \pm 0.7) \times 10^{-3}\ W(CO) - (8.0 \pm 1.5) \times 10^{-4}$$

$$I_{3000} = (4.9 \pm 0.1) \times 10^{-21}\ N(HI) + (2.5 \pm 0.5) \times W(CO) + (0.7 \pm 0.4)$$

14. G. Lagache, A. Abergel, F. Boulanger, J.-L. Puget, *Astron. Astrophys.* **333**, 709 (1998).
15. G. Lagache, *Astron. Astrophys.* **405**, 813 (2003).
16. A. W. Strong, J. R. Mattox, *Astron. Astrophys.* **308**, L21 (1996).
17. S. Digel, I. A. Grenier, S. D. Hunter, T. M. Dame, P. Thaddeus, *Astrophys. J.* **555**, 12 (2001).
18. P. Sreekumar *et al.*, *Astrophys. J.* **494**, 523 (1998).
19. S. D. Hunter *et al.*, *Astrophys. J.* **481**, 205 (1997).
20. A. W. Strong, I. V. Moskalenko, O. Reimer, *Astrophys. J.* **613**, 956 (2004).
21. C. Fichtel, R. G. Stone, M. E. Ozel, P. Sreekumar, *Astrophys. J.* **374**, 134 (1991).
22. M. Pohl, C. Perrot, I. Grenier, S. Digel, *Astron. Astrophys.* **409**, 581 (2003).
23. R. C. Bohlin, B. D. Savage, J. F. Drake, *Astrophys. J.* **224**, 132 (1978).
24. H. Yamamoto, T. Onishi, A. Mizuno, Y. Fukui, *Astrophys. J.* **592**, 217 (2003).
25. A. Mizuno *et al.*, *Publ. Astron. Soc. Jpn.* **53**, 971 (2001).
26. W. T. Reach, B.-C. Koo, C. Heiles, *Astrophys. J.* **429**, 672 (1994).
27. J. P. Williams, C. F. McKee, *Astrophys. J.* **476**, 166 (1997).
28. M. H. Heyer, J. M. Carpenter, R. L. Snell, *Astrophys. J.* **551**, 852 (2001).
29. F. Combes, *New Astron. Rev.* **46**, 755 (2002).
30. C. Gry *et al.*, *Astron. Astrophys.* **391**, 675 (2002).
31. L. Blitz, D. Bazel, F. X. Désert, *Astrophys. J.* **352**, L13 (1990).
32. C. Heiles, T. H. Troland, *Astrophys. J.* **586**, 1067 (2003).

33. We thank A. Strong, I. Moskalenko, and O. Reimer for providing their IC emission maps and the referees for their insightful remarks. Discussions with F. Boulanger, C. Heiles, J. L. Puget, M. Sauvage, and H. Aussel were extremely helpful. This research made use of the Legacy Archive for Microwave Background Data Analysis (LAMBDA) and High-Energy Astrophysics Science Archive Research Center (HEASARC) archives from NASA.

Supporting Online Material

www.sciencemag.org/cgi/content/full/307/5713/1292/DC1
Figs. S1 to S3
Table S1

29 October 2004; accepted 3 January 2005
10.1126/science.1106924

Heat Capacity of a Strongly Interacting Fermi Gas

Joseph Kinast,[1] Andrey Turlapov,[1] John E. Thomas,[1]* Qijin Chen,[2] Jelena Stajic,[2] Kathryn Levin[2]

We have measured the heat capacity of an optically trapped, strongly interacting Fermi gas of atoms. A precise addition of energy to the gas is followed by single-parameter thermometry, which determines the empirical temperature parameter of the gas cloud. Our measurements reveal a clear transition in the heat capacity. The energy and the spatial profile of the gas are computed using a theory of the crossover from Fermi to Bose superfluids at finite temperatures. The theory calibrates the empirical temperature parameter, yields excellent agreement with the data, and predicts the onset of superfluidity at the observed transition point.

Strongly interacting, degenerate atomic Fermi gases (*1*) provide a paradigm for strong interactions in nature (*2*). In all strongly interacting Fermi systems, the zero-energy scattering length is large compared with the interparticle spacing, a property that produces universal behavior (*3*, *4*). Predictions of universal interactions and effective field theories in nuclear matter (*3*, *5*–*7*) are tested by measurements of the interaction energy (*1*, *8*–*10*). Anisotropic expansion of strongly interacting Fermi gases (*1*) is analogous to the "elliptic flow" of a quark-gluon plasma (*2*). High-temperature superfluidity has been predicted (*11*–*16*) in strongly interacting Fermi gases, which can be used to test theories of high-temperature superconductivity (*17*). Microscopic evidence for superfluidity has been obtained by observing the pairing of fermionic atoms (*18*–*20*). Macroscopic evidence arises in anisotropic expansion (*1*) and in collective excitations (*21*–*23*).

In superconductivity and superfluidity, measurements of the heat capacity have played an important role in determining phase transitions (*24*) and in characterizing the nature of bosonic and fermionic excitations. We report on the measurement of the heat capacity for a strongly interacting Fermi gas of ^{6}Li atoms that are confined in an optical trap. Our experiments (*25*) examine the fundamental thermodynamics of the gas.

Thermodynamic properties of the Bardeen-Cooper-Schrieffer–Bose-Einstein condensation (BCS-BEC) crossover system are computed (*26*) using a consistent many-body theory (*27*, *28*) based on the conventional mean-field state (*29*). BCS-BEC crossover refers to the smooth change from the BCS superfluidity of fermions to the BEC of dimers, by varying the strength of the pairing interaction (for example, by tuning a magnetic field). The formalism of (*16*, *17*, *28*) was applied recently (*30*) to explain radio-frequency measurements of the gap (*20*). The theory contains two contributions to the entropy and energy arising from fermionic and bosonic excitations. The latter are associated principally with excited pairs of fermions (Cooper pairs at finite momentum). In this model, there is no direct boson-boson coupling, and fermion-boson interactions are responsible for the vanishing of the pair chemical potential μ_{pair} in the superfluid regions. The vanishing of μ_{pair} implies that, within a trap, the associated low-temperature power laws in the entropy and energy are the same as those of the homogeneous system (*31*). This is to be contrasted with models that involve noninteracting bosons and fermions (*32*). Our BCS-like ground state ansatz will be inapplicable at some point when the fermionic degrees of freedom have completely disappeared and the gas is deep in the BEC regime, where the power laws associated with true interacting bosons are expected (*31*). In that case, direct interboson interactions must be accounted for, and they will alter the collective-mode behavior (*33*). However, on the basis of collective-mode experiments (*21*–*23*) and their theoretical interpretation (*34*, *35*), one can argue that the BCS-like ground state appears to be appropriate in the near-resonance, unitary regime. The thermodynamic quantities within the trap are computed by using previously calculated profiles (*36*) of the various energy gaps and the particle density as a function of the radius.

Unlike the pairing gap in the weak-coupling BCS limit, the pairing gap in the unitary regime is very large. Well below the superfluid transition temperature T_c, fermions are paired over much of the trap, and unpaired fermions are present only at the edges of the trap. These unpaired fermions tend to dominate the thermodynamics associated with the fermionic degrees of freedom, and lead to a power law that is higher than linear in the temperature (T) dependence of the entropy. The contribution from finite-momentum Cooper pairs leads to a $T^{3/2}$ dependence of the entropy on temperature. Both bosonic and fermionic contributions are important at low T.

An important feature of these fermionic superfluids is that pair formation occurs at a higher temperature, T^*, than T_c, where pairs condense. At temperatures $T > T^*$, the entropy approaches that of the noninteracting gas. For $T_c < T < T^*$, the attraction is strong enough to form quasi-bound (or preformed) pairs, which are reflected in the thermodynamics. At these intermediate temperatures, a finite energy (i.e., the pseudogap) is needed to create single fermion excitations (*16*, *17*, *28*). In the unitary regime, both T^* and T_c are large fractions of the Fermi temperature T_F, signifying high-temperature pair formation and very high-temperature superfluidity.

We prepared a degenerate unitary Fermi gas composed of a 50:50 mixture of the two lowest spin states of ^{6}Li atoms near a Feshbach resonance. To cool the gas, we used forced evaporation at a bias magnetic field of 840 G in an ultrastable CO_2 laser trap (*1*, *2*, *26*). After cooling the atoms well into the degenerate regime, energy was precisely added to the trapped gas at fixed atom number. The gas was then allowed to thermalize for 0.1 s, and was released from the trap and imaged at 840 G after 1 ms of expansion to determine the number of atoms and the empirical temperature parameter $\tilde{T}$. For our trap, the total number of atoms was $N = 2.2\ (0.3) \times 10^5$. The corresponding noninteracting gas Fermi temperature was $T_F = (3N)^{1/3}\hbar\bar{\omega}/k_B \approx 2.5\ \mu K$, where $\hbar$ is Planck's constant divided by 2π, $\bar{\omega}$ is the geometric mean of the trap oscillation frequencies, k_B is Boltzmann's constant, and T_F is small compared $U_0/k_B = 35\ \mu K$, where U_0 is the final trap depth.

Energy was precisely added to the trapped gas at fixed atom number by releasing the cloud from the trap and permitting it to expand for a short time $0 \le t_{heat} \le 460\ \mu s$, after which the gas was recaptured. Even for a strongly interacting gas, the energy input is well defined for very low initial temperatures, where both the equation of state and the expansion dynamics are known. During the times t_{heat} used in these experiments, the axial size of the gas changed negligibly, whereas transverse dimensions expanded by a factor $b_\perp(t_{heat})$. Therefore, the mean harmonic-trapping potential energy $\langle U_{HO} \rangle$ in each of the two transverse directions increased by a factor $b_\perp^2(t_{heat})$.

The initial potential energy was readily determined at zero temperature from the

[1]Physics Department, Duke University, Durham, NC 27708–0305, USA. [2]James Franck Institute and Department of Physics, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA.

*To whom correspondence should be addressed. E-mail: jet@phy.duke.edu

equation of state of the gas, $(1 + \beta)\varepsilon_F(x) + U_{HO} = \mu_0$ (*1*, *8*), where $\varepsilon_F(x)$ is the local Fermi energy as a function of position, β is the unitary gas parameter (*1*, *3*, *6*–*8*), and μ_0 is the global chemical potential. This equation of state is supported by low-temperature studies of the breathing mode (*21*, *23*, *33*, *35*) and the spatial profiles (*1*, *6*, *36*). It is equivalent to that of a harmonically trapped noninteracting gas of particles with an effective mass (*5*), which in our notation is $m^* = m/(1 + \beta)$, where m is the bare-fermion mass. The mean potential energy is half of the total energy, because the gas behaves as a harmonic oscillator. Because $\beta < 0$ (*6*, *7*), $m^* > m$, so that the effective oscillation frequencies and the chemical potential are simply scaled down, yielding $\mu_0 = k_B T_F \sqrt{1 + \beta}$ (*1*, *8*). The total energy at zero temperature, which determines the energy scale, is therefore

$$E_0 = \frac{3}{4} N \mu_0 = \frac{3}{4} N k_B T_F \sqrt{1 + \beta} \quad (1)$$

For each direction, the initial potential energy at zero temperature is $E_0/6$. Then, the total energy of the gas after heating is given by

$$E(t_{heat}) = \eta E_0 \left[\frac{2}{3} + \frac{1}{3} b_\perp^2 (t_{heat}) \right] \quad (2)$$

neglecting trap anharmonicity (*26*). Here, η is a correction factor arising from the finite temperature of the gas before the energy input. For the strongly interacting gas, the initial reduced temperature is very low. We assume that it is $\tilde{T} \approx 0.04$, where $\tilde{T}$ is measured and calibrated as described below. Assuming a Sommerfeld correction then yields $\eta_{int} \approx 1 + 2\pi^2 \tilde{T}^2/3 \approx 1.01$, which hardly affects the energy scale.

A zero-temperature strongly interacting gas expands by a hydrodynamic scale factor $b_\perp^H (t_{heat})$, when released from a harmonic trap (*1*, *37*). Heating arises after recapture and subsequent equilibration, but not during expansion. This follows from the lowest $\tilde{T} = 0.04$, which is obtained by imaging the gas 1 ms after release from the trap. Hence, the temperature change during $t_{heat} \leq 460\ \mu s <$ 1 ms must be very small.

Thermometry of strongly interacting Fermi gases is not well understood. In contrast, thermometry of noninteracting Fermi gases can be simply accomplished by fitting the spatial distribution of the cloud (after release and ballistic expansion) with a Thomas-Fermi (T-F) profile, which is a function of two parameters. We choose the two parameters to be the Fermi radius σ_x and the reduced temperature T/T_F. However, this method is only precise at temperatures well below 0.5 T_F, where σ_x and T/T_F are determined independently. At higher temperatures, where the Maxwell-Boltzmann limit is approached, such a fit determines only the product $\sigma_x^2 \times T/T_F$. We circumvent this problem by determining σ_x from a low-temperature fit, and then hold it constant in the fits at all higher temperatures, which enables a one-parameter determination of the reduced temperature.

Spatial profiles of strongly interacting Fermi gases closely resemble T-F distributions, which were observed experimentally (*1*, *10*) and were predicted (*36*). The profiles of the trapped and released gas are related by hydrodynamic scaling to a good approximation. Over a wide temperature range, this scaling is consistent to ±2% with the observed cloud size and is further supported by measurements of the breathing frequency, which are within ±1% of the unitary hydrodynamic value (*21*). Analogous to the noninteracting case, we define an experimental dimensionless temperature parameter $\tilde{T}$, which is determined by fitting the cloud profiles with a T-F distribution holding constant the Fermi radius of the interacting gas, σ_x'. This method is inspired by the ideas presented in (*38*). We find experimentally that $\tilde{T}$ increases monotonically from the highly degenerate regime to the Maxwell-Boltzmann limit. This fitting procedure also leads us to define a natural reduced-temperature scale in terms of the zero-temperature parameters β and T_F,

$$\tilde{T}_{nat} \equiv \frac{k_B T}{\mu_0} = \frac{T}{T_F \sqrt{1 + \beta}} \quad (3)$$

Equation 3 is consistent with our choice of fixed Fermi radius σ_x', that is, $m\omega_x^2 \sigma_x'^2 / 2 = \mu_0$, where ω_x is the trap oscillation frequency in the x direction. At high temperatures, we must interpret $\tilde{T} = \tilde{T}_{nat}$ to obtain the correct Maxwell-Boltzmann limit. At low temperatures, $\tilde{T} \approx \tilde{T}_{nat}$ yields an estimate of T/T_F that can be further calibrated to the theoretical reduced temperature T/T_F by performing the experimental fitting procedure on the theoretically generated density profiles (*26*, *27*).

Preliminary data processing yields normalized, one-dimensional spatial profiles of the atomic cloud (*26*). To determine $\tilde{T}$ over the full temperature range of interest, we used a fixed expansion time of 1 ms. We first measured σ_x' from our lowest temperature data. Then, $\tilde{T}$ was determined using the one-parameter T-F fit method. This yielded $\tilde{T}$ from 0.04 to 2.15 for the strongly interacting gas.

The experimental energy scale Eq. 1 and the natural temperature scale Eq. 3 were determined by measuring the value of β. This was accomplished by comparing the measured radius of the strongly interacting gas, σ_x', to the radius for a noninteracting gas (*26*). We found that $\beta = -0.49$ (0.04) (statistical error only), which is in reasonable agreement with the best current predictions, where $\beta = -0.56$ (*6*) and $\beta = -0.545$ (*7*).

We next applied our energy input and thermometry methods to measure the heat capacity of our optically trapped Fermi gas. For different values of t_{heat}, we measured the temperature parameter $\tilde{T}$ and calculated the total energy $E(t_{heat})/E_0$ by using Eq. 2. The time t_{heat} determines the energy accurately because the trap intensity switches in less than 1 μs. We believe that shot-to-shot fluctuations in the energy are negligible, based on the small fractional fluctuations in $\tilde{T}$ at low temperatures, where the heat capacity is expected to be very small. To obtain high-resolution data, 30 to 40 different heating times t_{heat} were chosen. The data for each of these heating times were acquired in a random order to minimize systematic error. Ten complete runs were taken through the entire random sequence.

We first measured the heat capacity for a noninteracting Fermi gas (*21*, *26*), where the scattering length a was zero. This occurred near 526 G. Figure 1 shows the data (green dots) that represent the calculated $E(t_{heat})/E_0$ versus the measured value of $\tilde{T}$, for each t_{heat}. For comparison, predictions for a noninteracting trapped Fermi gas, where $E_{ideal}(T)/E_{ideal}(0)$, are shown as the black curve, where $\tilde{T} = T/T_F$ in this case. Here, the chemical potential and energy were calculated using a finite temperature Fermi distribution and the density of states for the trapped gas. Throughout, we used the density-of-states for a realistic Gaussian potential well, $U(r) = U_0[1 - \exp(-m\bar{\omega}^2 r^2 / 2U_0)]$, where $U_0 = 14.6 k_B T_F$ and r is the radial coordinate, rather than the harmonic oscillator approximation. This model is in very good agreement with the noninteracting gas data at all temperatures.

For the strongly interacting gas at 840 G [Fig. 1 (blue diamonds)], the gas was cooled to $\tilde{T} = 0.04$ and then heated. The temperature parameter $\tilde{T}$ varies by a factor of 50, and the total energy varies by a factor of 10. For comparison, we show the theoretical results for the unitary case as the red curve. Here, the horizontal axis for the theory was obtained using the approximation $\tilde{T} \approx \tilde{T}_{nat}$ via Eq. 3. On a large-scale plot, the data for the strongly interacting and noninteracting gases appear to be quite similar, although there are important differences at low temperature.

A noticeable result is observed by plotting the low-temperature data of Fig. 1 on an expanded scale (*25*, *26*). This reveals a transition in the heat capacity, which is made evident by plotting the data for the strongly interacting gas on a log-log scale, as in Fig. 2. The transition is apparent in the raw temperature data (*25*, *26*) and is strongly

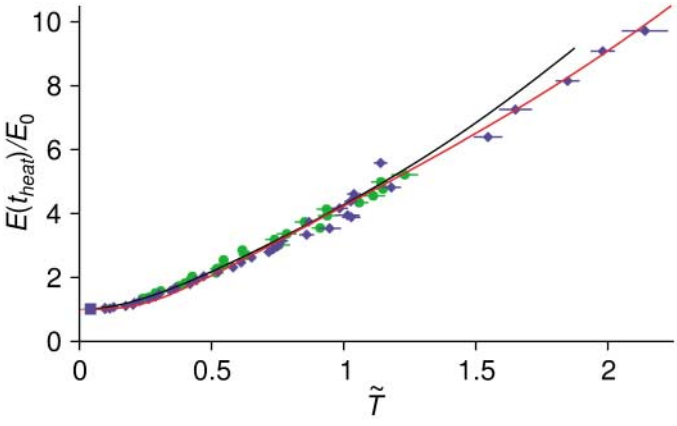


Fig. 1. Total energy versus temperature. For each heating time t_{heat}, the temperature parameter $\tilde{T}$ is measured from the cloud profile, and the total energy $E(t_{heat})$ is calculated from Eq. 2 in units of the ground state energy E_0. Green circles indicate noninteracting Fermi gas data and blue diamonds indicate strongly interacting Fermi gas data. The black curve shows the predicted energy versus reduced temperature for a noninteracting trapped Fermi gas, $E_{ideal}(\tilde{T})/E_{ideal}(0)$. The red curve shows the predicted energy versus $\tilde{T}$ for the unitary case. No temperature calibration is applied because $\tilde{T} \approx \tilde{T}_{nat}$ over the broad temperature range shown. Note that the lowest temperature point (blue square) is constrained to lie on the black curve.

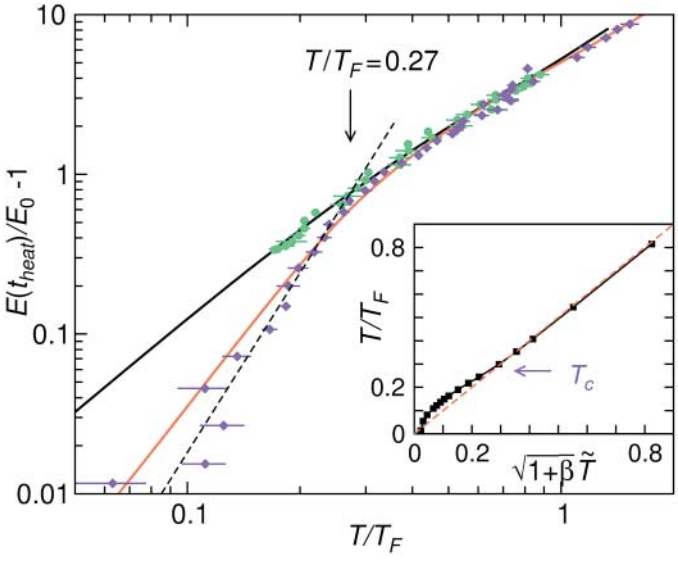


Fig. 2. Energy input versus temperature from Fig. 1 after temperature calibration on a log-log scale. The strongly interacting Fermi gas shows a transition in behavior near $T/T_F = 0.27$. Green circles indicate noninteracting Fermi gas data and blue diamonds indicate strongly interacting Fermi gas data. The red curve shows the prediction for a unitary Fermi gas in a Gaussian trap and the black curve shows the prediction for a noninteracting Fermi gas in a Gaussian trap, as in experiment. The black dashed line shows the best-fit power law $[97.3\ (T/T_F)^{3.73}]$ to the unitary data for $T/T_F \leq 0.27$. The inset shows the calibration curve, which has been applied to the unitary data (blue diamonds). The red dashed line in the inset represents the diagonal, $T/T_F = \sqrt{1+\beta}\tilde{T}$. Here, $E_0 \equiv E(T = 0)$.

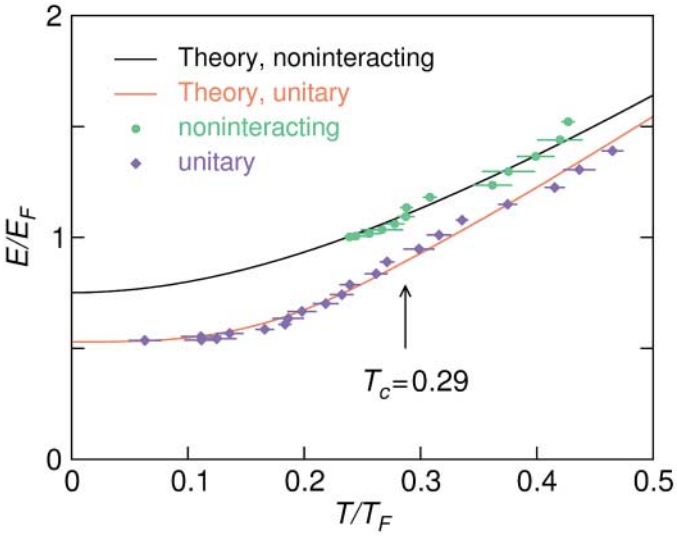


Fig. 3. Low-temperature comparison of present theory (red, black curves) and experiments (symbols) in terms of E/E_F ($E_F = k_B T_F$) per atom as a function of T/T_F, for both unitary and noninteracting gases in a Gaussian trap. The two experimental (and the two theoretical) curves do not merge until higher $T^* > T_c$, which is consistent with the presence of a pseudogap.

suggestive of the onset of superfluidity. The observed spatial profiles of the gas vary smoothly and are closely approximated by T-F shapes in the transition region. Figure 2 shows the transition after converting the empirical temperature $\tilde{T}$ to theoretical T/T_F units.

The empirical temperature was calibrated to enable precise comparison between the theory and the experimental data. For the calibration, we subjected the theoretically derived density profiles (*27*, *36*) to the same one-dimensional T-F fitting procedure that was used in the experiments. One-dimensional density distributions were obtained by integrating over two of the three dimensions of the predicted spatial profiles, which were determined for a spherically symmetric trap. Our results for this temperature calibration are shown in the inset to Fig. 2. This calibration provides a mapping between the experimental reduced temperature $\sqrt{1+\beta}\tilde{T}$, and the theoretical temperature T/T_F. We found that $\tilde{T} = \tilde{T}_{nat}$ is a very good approximation above T_c. Such scaling may be a manifestation of universal thermodynamics (*4*). The difference between $\tilde{T}$ and $\tilde{T}_{nat}$ is significant only below the superfluid transition T_c and is therefore negligible in the large-scale plot of Fig. 1 over a broad temperature range. However, below T_c, the fits to the theoretical profiles yield a value of $\sqrt{1+\beta}\tilde{T}$, which is lower than the theoretical value of T/T_F. This is a consequence of condensate effects (*26*).

Figure 2 shows that above a certain temperature T_c, the strongly interacting data nearly overlap that of the noninteracting gas, and exhibit a power law fit $E/E_0 - 1 = 4.98\ (T/T_F)^{1.43}$. Below T_c, the data deviate greatly from noninteracting Fermi gas behavior, and are well fit by $E/E_0 - 1 = 97.3\ (T/T_F)^{3.73}$ (dashed curve). From the intersection point of these power law fits, we estimate $T_c/T_F = 0.27$ (0.02) (statistical error only). This is very close to our theoretical value $T_c/T_F = 0.29$.

The fractional change in the heat capacity C is estimated from the slope change in the fits to the calibrated data. We find the relative specific heat jump $(C_< - C_>)/C_> \approx 1.51$ (0.05) (statistical error only), where $>$ denotes above T_c and $<$ denotes below T_c. This is close to the value (1.43) for an s-wave BCS superconductor in a homogeneous case, although one expects preformed pairs, i.e., pseudogap effects, to slight modify the discontinuity (*28*).

In Figs. 2 and 3, the theory is compared to the calibrated data after a very slight detuning of the magnetic field in the model away from resonance, so that the predicted unitary gas parameter β has the same value as measured. This small detuning, $(k_F a)^{-1} = 0.11$, where $k_F = \sqrt{2mk_B T_F/\hbar^2}$, is reasonable given the broad Feshbach resonance (*39*) in ^{6}Li.

Finally, Fig. 3 presents an expanded view of the low-temperature region. Here, the experimental unitary data are calibrated and replotted in the more conventional theoretical units, $E_F = k_B T_F$ and T_F. The agreement between theory and experiment is very good. In the presence of a pseudogap, a more elaborate treatment (*28*) of the pseudogap self energy, which takes into account spectral broadening, will be needed in order to calculate accurately the jump in specific heat.

By extending the temperature range in Fig. 3 to high T, we find that both the unitary and noninteracting cases coincide above a characteristic temperature, T^*, although below T_c they start out with different power laws (as shown in Fig. 2). In general, we find that agreement between theory and experiment is very good over the full temperature range for which the data were taken. The observation that the interacting and noninteracting curves do not precisely coincide until temperatures rise substantially above T_c is consistent with (although it does not prove) the existence of a pseudogap and with onset temperature from the figure $T \approx 2\ T_c$. Related signatures of pseudogap effects are also seen in the thermodynamics of high-temperature superconductors (*17*).

References and Notes

1. K. M. O'Hara, S. L. Hemmer, M. E. Gehm, S. R. Granade, J. E. Thomas, *Science* **298**, 2179 (2002).
2. J. E. Thomas, M. E. Gehm, *Am. Sci.* **92**, 238 (2004).
3. H. Heiselberg, *Phys. Rev. A.* **63**, 043606 (2001).
4. T.-L. Ho, *Phys. Rev. Lett.* **92**, 090402 (2004).
5. J. G. A. Baker, *Phys. Rev. C* **60**, 054311 (1999).
6. J. Carlson, S.-Y. Chang, V. R. Pandharipande, K. E. Schmidt, *Phys. Rev. Lett.* **91**, 050401 (2003).
7. A. Perali, P. Pieri, G. C. Strinati, *Phys. Rev. Lett.* **93**, 100404 (2004).
8. M. E. Gehm, S. L. Hemmer, S. R. Granade, K. M. O'Hara, J. E. Thomas, *Phys. Rev. A* **68**, 011401(R) (2003).
9. T. Bourdel *et al.*, *Phys. Rev. Lett.* **93**, 050401 (2004).
10. M. Bartenstein *et al.*, *Phys. Rev. Lett.* **92**, 120401 (2004).
11. M. Houbiers *et al.*, *Phys. Rev. A.* **56**, 4864 (1997).
12. R. Combescot, *Phys. Rev. Lett.* **83**, 3766 (1999).
13. M. Holland, S. J. J. M. F. Kokkelmans, M. L. Chiofalo, R. Walser, *Phys. Rev. Lett.* **87**, 120406 (2001).
14. E. Timmermans, K. Furuya, P. W. Milonni, A. K. Kerman, *Phys. Lett. A* **285**, 228 (2001).
15. Y. Ohashi, A. Griffin, *Phys. Rev. Lett.* **89**, 130402 (2002).
16. J. Stajic *et al.*, *Phys. Rev. A.* **69**, 063610 (2004).
17. Q. J. Chen, J. Stajic, S. Tan, K. Levin, available at http://arxiv.org/abs/cond-mat/0404274.
18. C. A. Regal, M. Greiner, D. S. Jin, *Phys. Rev. Lett.* **92**, 040403 (2004).
19. M. W. Zwierlein *et al.*, *Phys. Rev. Lett.* **92**, 120403 (2004).
20. C. Chin *et al.*, *Science* **305**, 1128 (2004).
21. J. Kinast, S. L. Hemmer, M. E. Gehm, A. Turlapov, J. E. Thomas, *Phys. Rev. Lett.* **92**, 150402 (2004).
22. M. Bartenstein *et al.*, *Phys. Rev. Lett.* **92**, 203201 (2004).
23. J. Kinast, A. Turlapov, J. E. Thomas, *Phys. Rev. A* **70**, 051401(R) (2004).
24. F. London, *Phys. Rev.* **54**, 947 (1938).
25. J. Kinast, A. Turlapov, J. E. Thomas, available at http://arxiv.org/abs/cond-mat/0409283.
26. Materials and methods are available as supporting material on *Science* Online.
27. Q. J. Chen, J. Stajic, K. Levin, available at http://arxiv.org/abs/cond-mat/0411090.
28. Q. J. Chen, K. Levin, I. Kosztin, *Phys. Rev. B* **63**, 184519 (2001).
29. A. J. Leggett, *Modern Trends in the Theory of Condensed Matter* (Springer-Verlag, Berlin, 1980), pp. 13–27.
30. J. Kinnunen, M. Rodríguez, P. Törmä, *Science* **305**, 1131 (2004).
31. L. D. Carr, G. V. Shlyapnikov, Y. Castin, *Phys. Rev. Lett.* **92**, 150404 (2004).
32. J. E. Williams, N. Nygaard, C. W. Clark, *N. J. Phys.* **6**, 123 (2004).
33. S. Stringari, *Europhys. Lett.* **65**, 749 (2004).
34. H. Hu, A. Minguzzi, X.-J. Liu, M. P. Tosi, *Phys. Rev. Lett.* **93**, 190403 (2004).
35. H. Heiselberg, *Phys. Rev. Lett.* **93**, 040402 (2004).
36. J. Stajic, Q. J. Chen, K. Levin, available at http://arxiv.org/abs/cond-mat/0408104.
37. C. Menotti, P. Pedri, S. Stringari, *Phys. Rev. Lett.* **89**, 250402 (2002).
38. B. Jackson, P. Pedri, S. Stringari, *Europhys. Lett.* **67**, 524 (2004).
39. M. Bartenstein *et al.*, available at http://arxiv.org/abs/cond-mat/0408673.

40. We thank T.-L. Ho, N. Nygaard, C. Chin, M. Zwierlein, M. Greiner, and D. S. Jin for stimulating correspondence. This research is supported by the Chemical Sciences, Geosciences, and Biosciences Division of the Office of Basic Energy Sciences, Office of Science, U.S. Department of Energy (DOE); the Physics Divisions of the Army Research Office; and NSF, the Fundamental Physics in Microgravity Research program of NASA, NSF-MRSEC grant DMR-0213745; and in part by the Institute for Theoretical Sciences, a joint institute of Notre Dame University and Argonne National Laboratory, and by the U.S. DOE, Office of Science through contract W-31-109-ENG-38.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1109220/DC1
Materials and Methods
Figs. S1 and S2
References and Notes

28 December 2004; accepted 16 January 2005
Published online 27 January 2005;
10.1126/science.1109220
Include this information when citing this paper.

Simultaneous State Measurement of Coupled Josephson Phase Qubits

R. McDermott,[1,2] R. W. Simmonds,[2] Matthias Steffen,[1] K. B. Cooper,[1] K. Cicak,[2] K. D. Osborn,[2] Seongshik Oh,[2] D. P. Pappas,[2] John M. Martinis[1]*

One of the many challenges of building a scalable quantum computer is single-shot measurement of all the quantum bits (qubits). We have used simultaneous single-shot measurement of coupled Josephson phase qubits to directly probe interaction of the qubits in the time domain. The concept of measurement crosstalk is introduced, and we show that its effects are minimized by careful adjustment of the timing of the measurements. We observe the antiphase oscillation of the two-qubit $|01\rangle$ and $|10\rangle$ states, consistent with quantum mechanical entanglement of these states, thereby opening the possibility for full characterization of multiqubit gates and elementary quantum algorithms.

Considerable progress has been made toward the implementation of a quantum computer (*1*) based on superconductors. Coherent single-qubit operations have been shown in Josephson flux (*2*) and phase (*3*) qubits, and the time domain interaction of coupled qubits (*4*) and a controlled-NOT logic gate (*5*) have been demonstrated in the Josephson charge qubit (*6*, *7*). Previous studies of coupled superconducting qubits have relied on separate measurements of the individual qubits (bitwise readout). Such an approach does not yield complete information about the system and fails, for example, to directly establish correlations between the qubits in the case of an entangled state. To test quantum algorithms efficiently or to perform quantum state tomography and thereby definitively prove entanglement, it is necessary to measure all the qubits simultaneously (wordwise readout) and with high fidelity. For multiqubit circuits with fixed coupling—a common architecture for superconducting qubits—the realization of this goal is complicated by measurement crosstalk: Measurement of the state of one qubit may perturb the state of other qubits, destroying information about quantum correlations. Although continued progress toward the realization of quantum gates in superconducting circuits requires a thorough understanding of measurement crosstalk, this issue has received little attention to date.

We describe simultaneous single-shot state measurements to probe the interaction of coupled Josephson phase qubits in the time domain. The observed antiphase oscillation of the occupation probabilities of the two-qubit basis states $|01\rangle$ and $|10\rangle$ is consistent with quantum mechanical entanglement of these

[1]Department of Physics, University of California, Santa Barbara, CA 93106, USA. [2]National Institute of Standards and Technology, 325 Broadway, Boulder, CO 80305, USA.

*To whom correspondence should be addressed. E-mail: martinis@physics.ucsb.edu

states. Moreover, the free evolution between the $|01\rangle$ and $|10\rangle$ states contains the essential ingredient of the two-qubit imaginary-SWAP (*i*-SWAP) operation, which, taken together with single-qubit rotations, forms a universal set of quantum gates (*8*). Our results depend crucially on a scheme for fast (~1 ns) state measurement and accurate adjustment of the timing of the measurements of the two qubits in order to circumvent measurement crosstalk in the circuit. The characteristic decay time for the two-qubit oscillations is consistent with the longitudinal relaxation time of the single-qubit circuit, which suggests that little additional dissipation is introduced by coupling the qubits. This bodes well for future tests of multiqubit gates and for more rigorous demonstrations of quantum correlations in multiqubit circuits.

The Josephson qubit can be thought of as a manufactured electrical "atom" having discrete energy levels that exist in a potential energy landscape determined by the circuit design parameters and bias (Fig. 1, A and B). We previously demonstrated high-resolution spectroscopy and coherent oscillations in the time domain in a single-qubit circuit (*9*, *10*). To implement a coupled qubit circuit, we connected two flux-biased phase qubits via a thin-film capacitor (*11*) (Fig. 1C). We label the qubits A and B. The interaction Hamiltonian can be written

$$H_{\mathrm{int}} = S/2\ (|01\rangle\langle 10| + |10\rangle\langle 01|) \qquad (1)$$

where $|01\rangle \equiv |0_A 1_B\rangle$. In terms of the circuit parameters, the coupling strength is $S \approx (C_x/C_j)\hbar\omega_{10}$, where C_x is the coupling capacitance, C_j is the junction self-capacitance, $\hbar$ is Planck's constant divided by 2π, and ω_{10} is the resonance frequency of the qubits. The interaction can be controlled by adjusting the flux bias of the qubits to change ω_{10}, bringing the qubits in and out of resonance. When the qubits are tuned to resonance (Fig. 2A, inset), the eigenstates are the (entangled) symmetric and antisymmetric combinations of $|01\rangle$ and $|10\rangle$, with eigenenergies $-S/2$ and $S/2$, respectively. Far from resonance, the system behaves as two independent qubits.

Because our circuit is a manufactured quantum system, the energy levels are not known a priori; therefore, we must use spectroscopy to map out the qubit resonance frequencies versus bias (*12*). In the frequency domain, the interaction is manifested as an avoided level crossing at the point where the resonance frequencies of the two qubits are matched (*13*). We biased qubit A to yield a resonance frequency $\omega_{10A}/2\pi$ of 8.65 GHz (*14*). Subsequent spectroscopy of qubit B revealed a splitting S/h = 80 MHz centered at 8.65 GHz (*15*). The measured splitting is consistent with the estimated coupling capacitance and junction self-capacitance of 6 fF and 700 fF, respectively, which are close to the design values.

We next investigated the interaction of the two qubits in the time domain. The qubits were tuned into resonance and initialized in the ground state $|00\rangle$. We then applied an 8.65-GHz microwave π pulse to qubit A to prepare the state $|10\rangle$ (Fig. 2A, inset). Because this state is not an eigenstate of the two-qubit Hamiltonian, it undergoes free evolution in the subspace spanned by the vectors $|01\rangle$ and $|10\rangle$, according to the relation

$$|\Psi(t)\rangle = \tfrac{1}{2}(|10\rangle + |01\rangle)\exp[(iS/2\hbar)t] + \tfrac{1}{2}(|10\rangle - |01\rangle)\exp[-(iS/2\hbar)t] = \cos[(S/2\hbar)t]|10\rangle + i\sin[(S/2\hbar)t]|01\rangle \qquad (2)$$

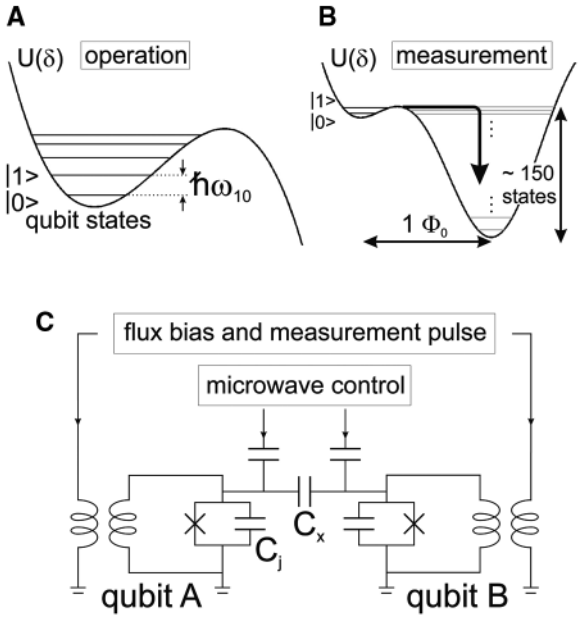


Fig. 1. (**A** and **B**) Potential energy landscape for the flux-biased Josephson phase qubit. During operation, the qubit is biased so that the junction phase δ is trapped in a metastable minimum of the potential $U(\delta)$, which contains several discrete energy levels. Measurement is accomplished by a fast flux pulse that adiabatically lowers the potential barrier, inducing a tunneling transition from the $|1\rangle$ state to the right-hand well of the potential, which contains around 150 states, resulting in a flux change of $\sim 1\Phi_0 \equiv h/2e$. (**C**) Circuit diagram of the coupled phase qubit circuit. The qubit junctions, with self-capacitance $C_j \approx 700$ fF, are coupled via the capacitance $C_x \approx 6$ fF; the junction critical currents are 1.7 μA, and the qubit loop inductances are 720 pH. The qubits are capacitively coupled to the microwave control lines; each qubit loop is inductively coupled to a coil that provides both the flux bias and the measurement pulse, as well as to a dc superconducting quantum interference device (SQUID; not shown) that is used to read out the flux state of the qubit loop. The devices are fabricated from Al/AlOx/Al trilayers by means of conventional thin-film techniques and optical lithography.

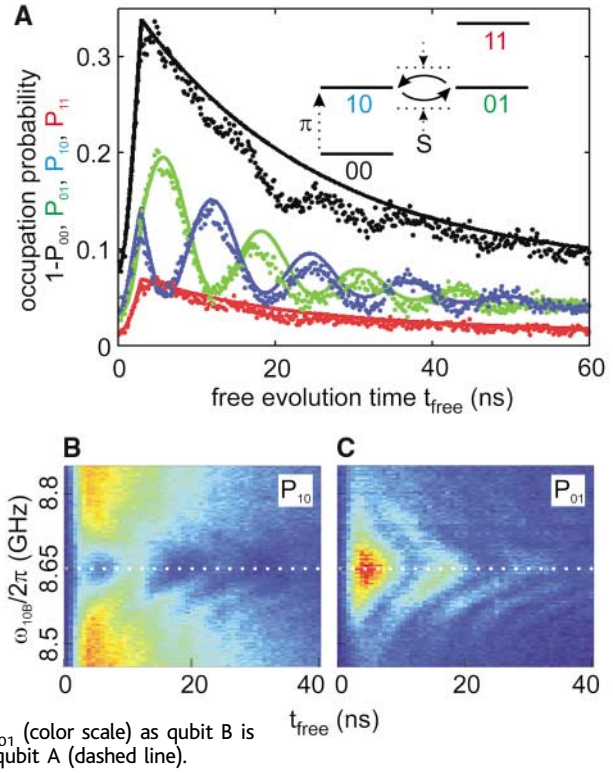


Fig. 2. Interaction of coupled qubits in the time domain. (**A**) With the qubits tuned to resonance $\omega_{10A}/2\pi = \omega_{10B}/2\pi$ = 8.65 GHz, a microwave π pulse on qubit A prepares the state $|10\rangle$. This state subsequently undergoes free precession in the subspace spanned by $|10\rangle$ and $|01\rangle$ before being measured, yielding the state $|00\rangle$, $|10\rangle$, $|01\rangle$, or $|11\rangle$. Repeated measurements give probabilities $1 - P_{00}$, P_{10}, P_{01}, and P_{11}, which are plotted (points) versus free precession time t_{free}. The solid lines are from numerical simulations that assume a gating time of 5 ns for the π pulse and take into account microwave cross-coupling of –10 dB, measurement fidelity of 70%, and a single-qubit T_1 of 25 ns, all determined experimentally from separate measurements. Inset shows the qubit energy levels and depicts schematically the coupling between the states $|10\rangle$ and $|01\rangle$. (**B** and **C**) Oscillations of P_{10} and P_{01} (color scale) as qubit B is detuned from the resonance of qubit A (dashed line).

Therefore, measurements of the two qubits should be anticorrelated, with a $|0\rangle$ for qubit A yielding a measurement of $|1\rangle$ for qubit B and a $|1\rangle$ for qubit A yielding a measurement of $|0\rangle$ for qubit B. After a variable period of free evolution t_{free}, we applied simultaneous measurement pulses to the two qubits, yielding four possible measurement outcomes. By repeated trials (100,000 events per data point), we obtained the occupation probabilities P_{00}, P_{01}, P_{10}, and P_{11}. These probabilities (points) are plotted in Fig. 2A versus t_{free}.

The occupation probabilities P_{01} and P_{10} oscillate out of phase, in agreement with the expected anticorrelation of the states of the two qubits. Moreover, the oscillation period is consistent with the 80-MHz splitting observed in the spectroscopy of the coupled qubits. The reduced amplitude of the oscillations is consistent with simulations (solid lines) that account for a gating time of 5 ns for the π pulse, a measured microwave cross-coupling from qubit A to qubit B of –10 dB, a measured 70% fidelity of the qubit state measurement (*10*), and a measured energy relaxation time T_1 of 25 ns for the individual qubits.

We next repeated the above experiment for a range of qubit detunings by adjusting the flux bias of qubit B. The occupation probabilities P_{01} and P_{10} as functions of both t_{free} and $\omega_{10B}/2\pi$ are shown in Fig. 2, B and C. The oscillation frequency increases and visibility decreases with detuning, resulting in a characteristic "chevron" pattern for the damped oscillations, in agreement with theoretical predictions (*10*).

We emphasize that the antiphase oscillation of the occupations of the $|01\rangle$ and $|10\rangle$ states is most clearly seen when the timing of the measurement pulses is adjusted to ensure simultaneity. As the relative delay t_{delay} of the measurement pulses is increased beyond ~2 ns, we observe a striking change in the character of the evolution of the occupation probabilities (Fig. 3). In the case of sequential measurements of the two qubits, measurement of the $|0\rangle$ state in the first qubit has no effect on the outcome of measurement of the second qubit. On the other hand, measurement of the $|1\rangle$ state in the first qubit results in an enhancement of the probability of measuring the $|1\rangle$ state in the second qubit. We refer to this phenomenon as measurement crosstalk.

The physical mechanism for measurement crosstalk in our circuit can be described as follows. The measurement of a $|1\rangle$ state in the first qubit implies a tunneling event to the right-hand well of the qubit potential. The resulting oscillation in the right-hand well produces a microwave voltage pulse (from the ac Josephson relation). This voltage drives a transient current $I_x(t)$ to the second qubit and induces transitions from the ground state. Because the qubits are weakly coupled ($C_x \ll C_j$), the effect of the current $I_x(t)$ can be understood by treating it as a classical drive to the second qubit. Numerical simulations indicate that the ring up of the second qubit can be separated into three segments in time (Fig. 4, A and B). In segment I, the initial oscillation in the right-hand well of the measured qubit samples the region near the turning point at the top of the well, corresponding to frequencies below

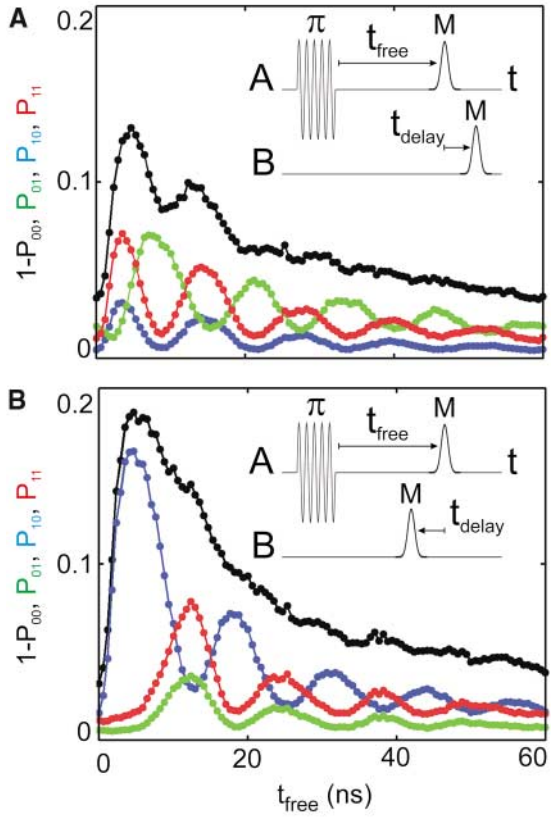


Fig. 3. Measurement crosstalk in the phase qubit, determined by sequential measurement of the states of the qubits. (**A**) Qubit A was measured 4 ns before qubit B; the data are plotted as in Fig. 2. (**B**) Same as (A) but with qubit A measured 4 ns after qubit B. The insets represent the currents applied to each of the qubit junctions; the microwave π pulse prepares the state $|10\rangle$ and Gaussian pulses (labeled *M*) measure the qubit states. When qubit A is measured before qubit B, the oscillations in P_{11} are correlated with the oscillations in P_{10}; when qubit A is measured after qubit B, the oscillations in P_{11} are correlated with the oscillations in P_{01}. From the relative amplitude of the oscillations, we conclude that measurement of $|1\rangle$ in the first qubit results in false measurement of $|1\rangle$ in the second qubit with ~70% probability.

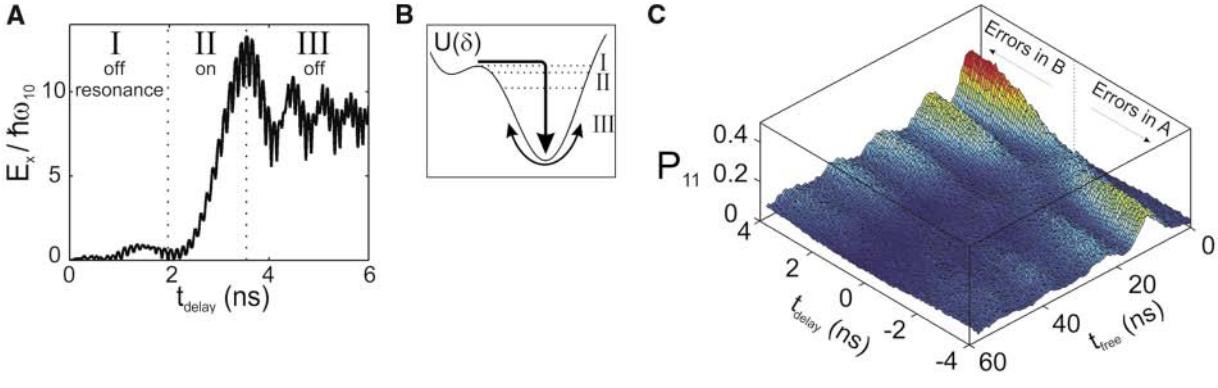


Fig. 4. Description of simultaneous measurement. (**A**) Numerical simulation of the energy transfer to qubit B induced by a tunneling event in qubit A. (**B**) The three stages of energy transfer correspond to energy decay through regions I to III in the potential diagram. (**C**) P_{11} versus free evolution time t_{free} and relative delay t_{delay} of the measurements of the two qubits.

the resonance frequency of the second qubit. The resulting drive current is off resonance with the second qubit; therefore, no appreciable coupling occurs for a time $\sim 0.1T_1$. In segment II, as the oscillations damp and come into resonance with the second qubit, the energy transferred to the second qubit is roughly quadratic in time and can be parameterized as $E_x/\hbar\omega_{10} \sim 10(C_x/C_j)^2[\omega_{10}(t_{delay} - 0.1T_1)]^2$. In segment III, as the measured qubit continues to decay and begins to sample the deepest harmonic regions of the right-hand well, the oscillation frequency moves above the resonance frequency of the second qubit. No additional energy is added, and the energy transferred to the second qubit levels out at a value $E_x/\hbar\omega_{10} \sim 100(C_x/C_j)^2\omega_{10}T_1$. Taking the probability for an $|0\rangle \rightarrow |1\rangle$ transition to be $P_1 \approx E_x/\hbar\omega_{10}$ for $E_x/\hbar\omega_{10} \ll 1$, we predict minimal measurement crosstalk for our circuit for $|t_{delay}| <$ 2 ns. Moreover, we note that the constraint on measurement timing becomes less stringent for qubits with longer T_1.

We investigated the dependence of measurement crosstalk on the timing of the measurements by repeating the experiment of Fig. 2 while varying t_{delay} to cover a total range of ±4 ns (Fig. 4C). When $t_{delay} > 2$ ns, the probability P_{11} is correlated with P_{10}; when $t_{delay} < 2$ ns, P_{11} is correlated with P_{01}. It is only when the relative delay of the measurements is optimally adjusted ($|t_{delay}| <$ 2 ns) that P_{11} is small and the oscillations in P_{11} disappear. Separate experiments indicate that when the timing of the measurement pulses is optimized, a tunneling event in one qubit results in a false measurement of $|1\rangle$ in the second qubit with only 15% probability. This residual measurement crosstalk can be attributed to the finite duration of the measurement pulse.

Our results suggest that it is possible in principle to perform high-fidelity measurements of multiple qubits. Such a technique may lead to scalable quantum information processing based on Josephson junctions.

References and Notes

1. M. A. Nielsen, I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge Univ. Press, Cambridge, 2000).
2. I. Chiorescu, Y. Nakamura, C. J. P. M. Harmans, J. E. Mooij, *Science* **299**, 1869 (2003); published online 13 February 2003 (10.1126/science.1081045).
3. J. M. Martinis, S. Nam, J. Aumentado, C. Urbina, *Phys. Rev. Lett.* **89**, 117901 (2002).
4. Yu. A. Pashkin *et al.*, *Nature* **421**, 823 (2003).
5. T. Yamamoto, Yu. A. Pashkin, O. Astafiev, Y. Nakamura, J. S. Tsai, *Nature* **425**, 941 (2003).
6. Y. Nakamura, Yu. A. Pashkin, J. S. Tsai, *Nature* **398**, 786 (1999).
7. D. Vion *et al.*, *Science* **296**, 886 (2002).
8. J. Kempe, K. B. Whaley, *Phys. Rev. A* **65**, 052330 (2002).
9. R. W. Simmonds, K. M. Lang, D. A. Hite, D. P. Pappas, J. M. Martinis, *Phys. Rev. Lett.* **93**, 077003 (2004).
10. K. B. Cooper *et al.*, *Phys. Rev. Lett.* **93**, 180401 (2004).
11. P. R. Johnson *et al.*, *Phys. Rev. B* **67**, 020509 (2003).
12. J. M. Martinis, M. H. Devoret, J. Clarke, *Phys. Rev. Lett.* **55**, 1543 (1985).
13. A. J. Berkley *et al.*, *Science* **300**, 1548 (2003); published online 15 May 2003 (10.1126/science.1084528).
14. At this frequency, the spectra of both qubits were free of fine structure normally attributed to spurious junction resonances (*9*).
15. R. McDermott *et al.*, data not shown.

16. Supported in part by the National Security Agency (NSA) Advanced Research and Development Activity (ARDA) through Army Research Office grants W911NF-04-1-2004 and MOD717304.

16 November 2004; accepted 6 January 2005
10.1126/science.1107572

Evidence for a Great Medieval Earthquake (~1100 A.D.) in the Central Himalayas, Nepal

J. Lavé,[1]*† D. Yule,[2]* S. Sapkota,[3] K. Basant,[3] C. Madden,[4] M. Attal,[1] R. Pandey[3]

The Himalayan orogen has produced three thrust earthquakes with moment magnitude (M_w) 7.8 to 8.5 during the past century, yet no surface ruptures associated with these great earthquakes have been documented. Here, we present paleoseismic evidence from east central Nepal that, since ~700 A.D., a single earthquake ruptured the Frontal Thrust fault at ~1100 A.D., with a surface displacement of ~17 (+5/–3) meters and a lateral extent and size that could have exceeded 240 kilometers and ~M_w 8.8, respectively. Ruptures associated with M_w <8.2 events would contribute to the frontal Himalayas folding but would stop before reaching the surface. These findings could require substantial modifications to current regional seismic hazard models.

The primary features of the Himalayan orogen are now understood, but the details of its seismotectonic behavior and maximum earthquake magnitudes are mostly unknown, despite their important implications regarding the seismic hazards facing densely populated regions. During the past century, the Himalayan arc has experienced three major thrust earthquakes of moment magnitude (M_w) >7.8. Growth folding (*1*, *2*) and surface faulting (*3*–*5*) have been reported in Holocene strata and terraces; paradoxically, none of these recent events reportedly produced coseismic surface ruptures, including the 1934 Bihar Nepal M_w 8.1 earthquake (*6*), which produced high-intensity shaking that was experienced throughout east Nepal and bordering regions of India (Fig. 1). To confirm the absence of rupture associated with this event and determine which events have led to the tectonic scarps, we conducted a paleoseismic study across the Himalayan front in the Marha Khola region, southeast of Kathmandu, in an area close to the inferred 1934 rupture zone.

Since ~20 million years ago, the deformation front resulting from the India/Asia collision has been expressed through the activation of two major thrust zones that are presumed to branch upward from a major midcrustal decollement: the Main Himalayan Thrust (*7*–*10*) (Fig. 1). In front of the rising Himalayas, thin-skinned thrust faulting has incorporated Cenozoic molasse deposits (Siwalik Formations) into the hanging walls of thrust faults, now expressed as the low-relief Siwalik Hills at the southern edge of the range. Geomorphic evidence of active tectonics indicates that 50 to 100% of the shortening across the Himalayas is transferred toward the southernmost of these faults, the Main Frontal Thrust fault (*2*, *5*, *11*). However, geodetic observations (*12*, *13*) indicate that current interseismic deformation is centered on a belt of microseismicity (*9*) that follows the southern edge of the Tibetan Plateau, ~100 km north of the frontal structure (Fig. 1). One explanation of this apparent paradox proposes that the current deformation mostly accumulates elastically at the transition along the decollement from steady creep beneath southern Tibet to locked beneath the High Himalaya, and that this elastic deformation releases and transfers to the front during large earthquakes (*2*, *9*), possibly like the 1934 M_w 8.1 earthquake.

[1]Laboratoire de Géodynamique des Chaînes Alpines, BP53, 38041 Grenoble, France. [2]California State University, Northridge, CA 91330, USA. [3]Seismolab, Department of Mines and Geology, Lainchaur, Kathmandu, Nepal. [4]Earth Consultants International, Tustin, CA 92780, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: jlave@ujf-grenoble.fr

The study area at Marha Khola is well suited to test the above seismotectonic model: It is located west of the maximum intensity felt during this 1934 event (Fig. 1) but within the hypothetical 200- to 300-km-wide rupture segment (*14*) and also across the active Bagmati/Ratu anticline (fig. S1) (*15*), where numerous folded fluvial Holocene terraces indicate a full transfer of the convergence to the frontal structure (*2*). Along the Marha Khola, fluvial strath terraces (Fig. 2) have been uplifted 5 to 40 m above its present channel. A tectonic scarp marks the southern extent of these terraces (fig. S1) (*15*). The scarp is ~4 m high across the youngest uplifted terrace (HT_4) and exhibits a sinuous geometry (Fig. 2).

Trenches 1 and 2, excavated across this scarp, and trench 3, a riverbank exposure cleaned and extended to depth (Fig. 2), expose four stratigraphic units (Fig. 3). Unit 1, deposited above the HT_4 strath at ~700 A.D. (samples MA-4 or M-2-29), consists of a near-uniform, 3-m-thick sequence of fluvial gravel (Fig. 2B). Unit 2 overlies this sequence and varies laterally from a 20-cm-thick, organic-rich soil (trench 3) dated at 1000 to 1200 A.D. (samples M-3-5 and M-2-9) to >2.5 m of fine to coarse-grained fan material, with interlayered thin soil horizons (trench 1). Fluvial incision has truncated units 1 and 2 along a major unconformity, U_{2-3}. In trench 2, the unconformity is overlain in its western part by a sequence of fluvial gravels (unit 3) that bury the footwall and deposit a thin veneer on top of the frontal fold. Subsequently, unit 4 deposits (interlayered weak soil, slope wash, overbank, and fluvial materials capped by the present-day 40-cm-thick cultivated soil) were deposited between 1000 and 1900 A.D. (Fig. 4).

The tectonic scarp consists of several main fault zones (F1, F2, and F3) and related folds. All faults in trenches 1 and 2 and F3 in trench 3 have ruptured the fluvial sequence (units 1 and 2) and are sealed by units 3 and 4. The upper ruptures F1 and F2 in trench 3 break units 1 and 2, but no dateable constraint in the form of an intact overlying horizon exists (Fig. 3). According to the charcoal ages (Fig. 4 and table S1) (*15*), the faults ruptured the surface between 1020 and 1160 A.D. Although a complex scenario with two large ruptures spanning ~150 years could be envisaged, the simplest explanation of our observations, supported by a careful analysis of relationships in T2 (discussion S1) (*15*), is to conclude that faults F1, F2, and F3 are coeval and the timing of their rupture is constrained by the age of formation of the major U_{2-3} unconformity at ~1100 A.D. (Fig. 4).

Along the right bank of Marha Khola and northwest of the scarp, the HT_4 strath level and overlying unit 1 gravels exhibit a near-uniform uplift relative to the present Marha Khola floodplain (Fig. 2), consistent with a rigid-body translation above a thrust ramp at depth (*2*). Assuming an ~S20°W slip direction (Fig. 1) and uniform rigid-body uplift above a ~N70°W-striking 25° ± 5° dipping ramp (fig. S1) (*15*), the 7- to 7.5-m vertical separation of HT_4 would result from a slip event of 17 (+6/–4) m. In the trenches, a similar amount of slip, 17 (+5/–3) m, can be estimated from the strike and dip of the different fault segments and from the vertical separation, at some distance from the fault, of either the HT_4 strath level or the top of the fluvial sequence (unit 1) (table S2 and discussion S2) (*15*).

According to the classical view on scaling between slip value, rupture area, and magnitude (*16*), we would expect that the ~1100 A.D. earthquake ruptured a large segment of the Himalayan arc. Unfortunately, published scaling theories are inappropriate to estimate the extent and magnitude of large shallow thrust earthquakes (*16*). Consequently, the surface rupture observed at Marha Khola suggests two possible end-member hypotheses: a large earthquake ($7.6 \leq M_w \leq 8.1$) with slip enhancement close to the surface or a great earthquake ($8.4 \leq M_w \leq 8.8$) activating a large fault plane with an average rupture slip similar to the value observed at Marha Khola.

A possible analog to the first hypothesis, the recent M_w 7.6 Chi-Chi thrust earthquake (*17*, *18*) in the western foothills of Central Taiwan, locally produced surface slip of >10 m,

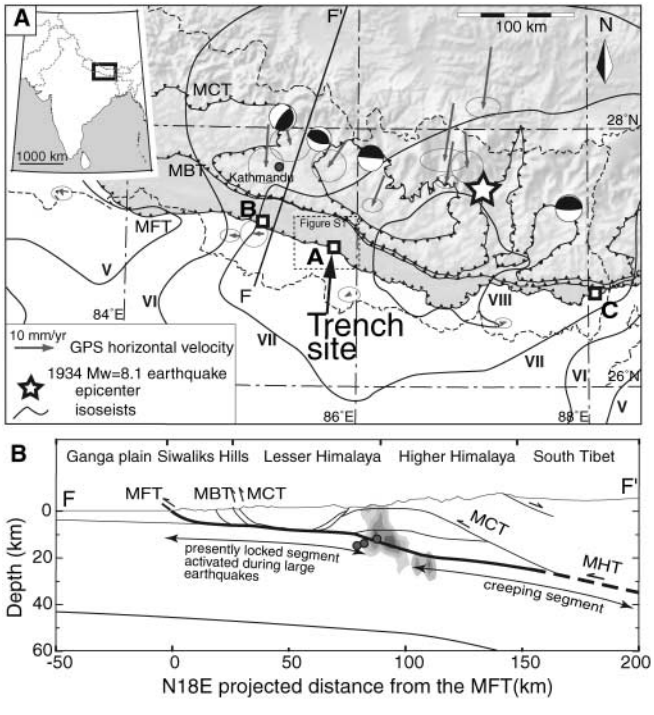


Fig. 1. (**A**) Map of eastern Nepal, including Global Positioning System (GPS) velocities relative to GPS stations in the Gangetic Plain (*23*), the focal mechanisms of the major earthquakes (M_w > 5) since 1965 [International Seismological Centre; Harvard solution (*24*)], and the epicenter of the M_w 8.1 1934 thrust earthquake (star) (*25*). The Marha Khola trench site (A) is located west of the most heavily shaken area (isoseismal MSK intensity contour = VIII) (*26*), ~240 km from a previous paleoseismic trenching study (C) in far east Nepal (*1*). At B, the Main Frontal Thrust fault slips at 21 ± 1.5 mm/year during the Holocene (*2*). (**B**) Simplified structural cross section across the central Himalayas of Nepal, with the major instrumental thrust earthquakes since 1965 (circles) and the cluster of microseismicity (*9*) (gray shading). MCT, Main Central Thrust; MBT, Main Boundary Thrust; MFT, Main Frontal Thrust; MHT, Main Himalayan Thrust.

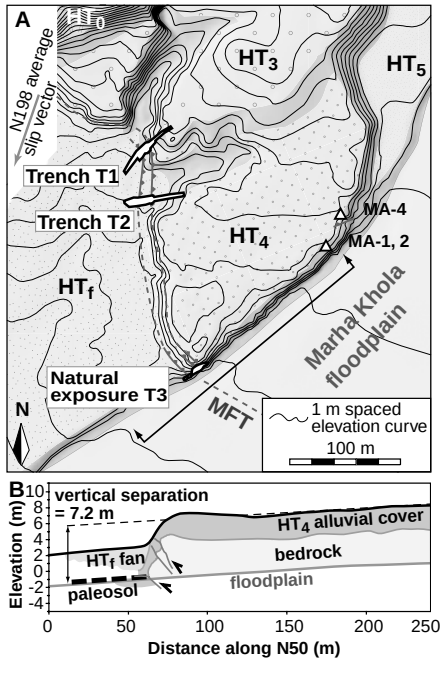


Fig. 2. (**A**) Topographic map of the trench site showing the location of trench exposures and the tectonic scarp of the Main Frontal Thrust fault (MFT) and uplifted terraces on the right bank of Marha Khola. (**B**) Cross section shows vertical separation across the fault of the top of HT_4 alluvial cover (vertical exaggeration = 5:1). Since scarp formation, active sedimentation (HT_f fan deposition) has occurred on the footwall to the southwest of the fault.

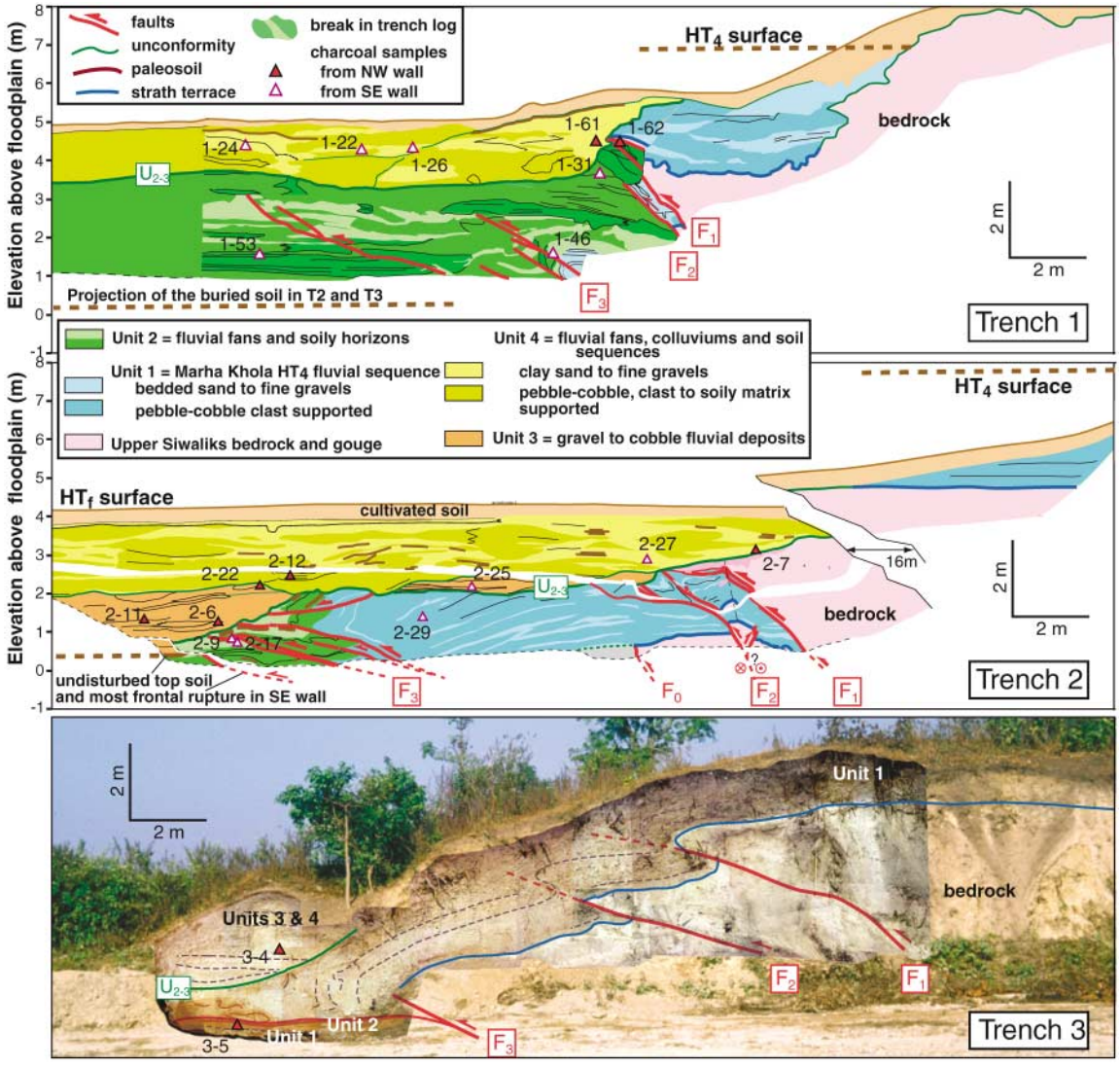


Fig. 3. Logs of the northwest walls of trenches 1 and 2 (**top**) and photograph of the northwest wall of trench 3 (**bottom**). Unconformity U_{2-3} (heavy green line) separates deformed from undeformed units, separating underlying pre-rupture from overlying post-rupture units, respectively. Triangles show locations of charcoal samples (see Fig. 4 for age data).

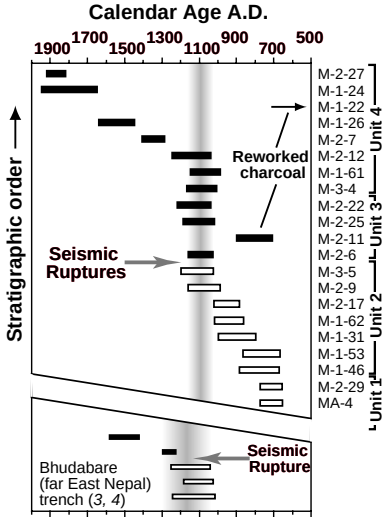


Fig. 4. Summary of the calendar radiocarbon ages, calibrated using CALIB 4.3 (*27*), from Marha Khola's trenches (table S1) (*15*) and from a trench in far East Nepal (*3*, *4*).

yet activated a relatively small rupture plane (30 × 80 km). Here, a Chi-Chi–type earthquake, of magnitude $M_w \leq 7.8$ (same rupture area but doubling slip), would release tectonic loading accumulated below the Siwaliks Hills or the southern part of the Lesser Himalayas.

A major conclusion of this study is the absence of surface rupture during the M_w 8.1 1934 Bihar Nepal earthquake, confirming previous reports (*3*, *14*). The rupture could have broken the vertical northern branch of the fault propagation fold (fig. S1) (*15*), but we did not observe a tectonic scarp across this northern fault, and subsidence in the Gangetic plain detected by leveling measurements (*19*) contradicts the hypothesis that the 1934 rupture propagated south of the Main Frontal Thrust fault (*7*). We suspect that earthquakes such as the 1934 event therefore could break the locked segment of the Main Himalayan Thrust (*2*, *12*) but die out before reaching the surface and deliver elastic strain at the southern tip of their rupture to active folding in the Siwalik Hills. This deformation, if elastically stored, could then be released during a Chi-Chi–type earthquake that activates the ramp and ruptures to the surface trace of the Main Frontal Thrust. Alternatively, a subsequent 1934-type event could activate the elastically stored energy and finally deliver slip to the surface, resulting in an unusually large surface displacement (maximized near the surface).

Two observations do not seem, however, to support these views. First, a summation of the seismic moment for the Himalayan arc reveals that the frequency of $M_w \cong 8.1$ earthquakes during the past three centuries is insufficient (*19*) to explain the transfer of most of the South Tibet/India convergence toward the frontal fold (*2*). Second, the absence of a surface rupture since the ~1100 A.D. event requires that elastic strain has been accumulating for ~830 years

beneath the Siwaliks Hills. However, the uplift profiles of the youngest Holocene terraces across the Bagmati/Ratu anticline (fig. S1) (*15*) do not exhibit the expected component of elastic deformation (discussion S3 and fig. S2) (*15*). Rather, terrace profiles suggest that the ruptures associated with blind M_w <8.2 events contribute by permanent postseismic deformation to the long-term folding of the most frontal Himalayan structures.

In the second hypothesis, the large slip value observed at Marha Khola would have been produced by a great earthquake that ruptured a large segment of the Himalayan arc. In contrast with the M_w = 8.1 1934-type events, its seismic energy would have been sufficient for the rupture to propagate up to the surface. In this hypothesis, the ~1100 A.D. earthquake may have nucleated below the High Himalayas, as is suspected for the 1934 event, and broken through to the surface trace of the Main Frontal Thrust fault, a cross-strike distance of about 100 km (Fig. 1). Its lateral extent could have therefore reached or overcome the 200- to 300-km length that has been ascribed to the 1934 M_w 8.1 earthquake (*14*). A trench across the Main Frontal Thrust fault in far east Nepal (*3*, *4*) exposed a rupture with >4 to 8 m of slip that occurred between 1050 and 1300 A.D. (Fig. 4). If this surface rupture is synchronous to the ~1100 A.D. event, its lateral extent would reach at minimum 240 km. The magnitude of this earthquake would range from M_w 8.4 up to M_w 8.8 (*20*), assuming that the respective slip values observed in far east Nepal and Marha Khola trenches are representative of the average slip on the fault plane.

Additional paleoseismic studies are required to test the above hypotheses and to ascribe a more precise rupture length, mean slip, and magnitude to the inferred great medieval earthquake. For the moment, according to the worst scenario and to magnitude distribution law for thrust earthquakes in Nepal (*21*), the return period for a ~M_w 8.8 and 17-m mean slip event would range between 1800 and 3000 years at Marha Khola (with large uncertainties, given that the magnitude distribution law can not easily be extrapolated to large earthquakes). Such very large events would thus accommodate 25 to 50% of the shortening across the Himalayas, and if generalized to the whole Himalayan arc, would help to bring the seismic moment summation to closure (*19*). Between two such great earthquakes, the Main Himalayan Thrust can generate, in addition, several $7.5 < M_w < 8.2$ events, but in the absence of surface ruptures, estimating the recurrence and slip for this type of Himalayan thrust earthquake is not possible using conventional paleoseismic trench studies.

References and Notes

1. T. Nakata, *Spec. Pap. Geol. Soc. Am.* **232**, 243 (1989).
2. J. Lavé, J. P. Avouac, *J. Geophys. Res.* **105**, 5735 (2000).
3. T. Nakata *et al.*, *Eos* **79**, F615 (1998).
4. B. N. Upreti *et al.*, Eds., *Proceedings of the Hokudan International Symposium and School on Active Faulting* (Letter Press, Hiroshima, Japan, 2000), pp. 533–536.
5. S. Kumar *et al.*, *Science* **294**, 2328 (2001).
6. N. N. Ambraseys, J. Douglas, *Geophys. J. Int.* **159**, 165 (2004).
7. L. Seeber, J. Armbuster, Eds., *Earthquake Prediction: An International Review* (Maurice Ewing Series, American Geophysical Union, Washington, DC, 1981), pp. 259–277.
8. W. Zhao, K. D. Nelson, Project INDEPTH Team, *Nature* **366**, 557 (1993).
9. M. R. Pandey, R. P. Tandukar, J. P. Avouac, J. Lavé, J. P. Massot, *Geophys. Res. Lett.* **22**, 751 (1995).
10. J. Lavé, J. P. Avouac, *J. Geophys. Res.* **106**, 26,561 (2001).
11. S. G. Wesnousky, S. Kumar, R. Mohindra, V. C. Thakur, *Tectonics* **18**, 967 (1999).
12. M. Jackson, R. Bilham, *J. Geophys. Res.* **99**, 13897 (1994).
13. R. Bilham, K. Larson, J. Freymuller, Project Idylhim Members, *Nature* **386**, 61 (1997).
14. M. R. Pandey, P. Molnar, *J. Geol. Soc. Nepal* **5**, 22 (1988).
15. Supporting material is available on *Science* Online.
16. C. H. Scholz, *The Mechanics of Earthquakes and Faulting* (Cambridge Univ. Press, Cambridge, ed. 2, 2002), pp. 179–243.
17. H. F. Ma, *Eos* **80**, 605 (1999).
18. S. Dominguez, J. P. Avouac, R. Michel, *J. Geophys. Res.* **108**, 2083 (2003).
19. R. Bilham, V. Gaur, P. Molnar, *Science* **293**, 1442 (2001).
20. M_w was computed from Kanamori's relation (*22*) for the seismic moment $M_w = 2/3\log_{10}(M_0) - 6$, where $M_0 = \mu \times \text{slip} \times L \times W$ (where L and W are the rupture length and width), the shear modulus $\mu = 3.3 \times 10^{10}$ N/m^2, $W = 100$ km from the locking zone below the High Himalayas to the Main Frontal Thrust fault, $L \geq 240$ km, and the average slip equals 5 to 17 m.
21. J. P. Avouac, L. Bollinger, J. Lavé, R. Cattin, M. Flouzat, *C. R. Acad. Sci.* **333**, 513 (2001).
22. H. Kanamori, *Tectonophysics* **93**, 185 (1983).
23. K. Larson, R. Bürgmann, R. Bilham, J. Freymueller, *J. Geophys. Res.* **104**, 1077 (1999).
24. P. Molnar, *J. Him. Geol.* **1**, 131 (1990).
25. W. P. Chen, P. Molnar, *J. Geophys. Res.* **82**, 2945 (1977).
26. N. N. Ambraseys, J. Jackson, *Curr. Sci.* **84**, 570 (2003).
27. M. Stuiver, P. J. Reimer, *Radiocarbon* **35**, 215 (1993).

28. This research was supported by the French program Programme National de Recherche sur les Risques Naturels. We are most grateful to the National Seismological Center, Kathmandu, and the Laboratoire de Détection Géophysique for their logistic help in the organization of the field surveys. A. Gajurel kindly provided help logging the trenches. We are indebted to G. Seitz for analyzing some of the charcoal samples at the Center for Accelerator Mass Spectrometry, Lawrence Livermore National Laboratory, California. We thank J. P. Avouac, K. Sieh, and two anonymous reviewers for insightful comments on earlier drafts of the manuscript.

Supporting Online Material

www.sciencemag.org/cgi/content/full/307/5713/1302/DC1
SOM Text
Figs. S1 and S2
Tables S1 and S2
References

2 September 2004; accepted 18 January 2005
10.1126/science.1104804

How Science Survived: Medieval Manuscripts' "Demography" and Classic Texts' Extinction

John L. Cisne

Determining what fraction of texts and manuscripts have survived from Antiquity and the Middle Ages has been highly problematic. Analyzing the transmission of texts as the "paleodemography" of their manuscripts yields definite and surprisingly high estimates. Parchment copies of the foremost medieval textbooks on arithmetical and calendrical calculation closely fit age distributions expected for populations with logistic growth and manuscripts with exponential survivorship. The estimated half-lives of copies agree with Bischoff's paleographically based suggestion that roughly one in seven manuscripts survive in some form from ninth-century Carolingian workshops. On this basis, many if not most of the leading technical titles circulating in Latin probably survived, even from late Antiquity.

Every student learns that the germ of his or her science barely made it through the Middle Ages. Just how likely were individual handwritten books to survive, or entire works to be lost? So far, the best evidence has come from the histories of individual manuscripts, libraries, and texts as reconstructed by paleographers, and it is overwhelmingly anecdotal.

Because manuscripts, unlike printed material, must be copied individually from antecedents, like organisms, their multiplication likewise should be inherently exponential. By treating the manuscripts of a

Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, NY 14853, USA. E-mail: cisne@geology.cornell.edu

text as if they were fossils from an extinct population (*1*, *2*), and by drawing on centuries of paleographic scholarship for demographic data, it becomes possible to roughly characterize the dynamics of the transmission process.

The Markov birth-and-death process is perhaps the simplest and most generally applicable stochastic model for any such statistical population's growth (*3*). Suppose that, in any given instant, each manuscript of a text has a certain probability λ of "giving birth" (being copied) and a certain probability μ of "death" (being destroyed). The population's expected per capita birthrate will be λ, and its expected per capita death rate will be μ. Individuals will die off exponentially with half-life (ln 2)/μ. If λ > μ, a population of size M_t at time t will increase exponentially with doubling time (ln 2)/(λ − μ) and will suffer chance extinction with probability $(\mu/\lambda)^{M_0}$ (*3*).

Of course, any real population's per capita growth rate will decrease as M_t increases toward some limit M_∞. Suppose that the per capita birth rate λ_M decreases linearly with the difference between the population's actual and limiting sizes: $\lambda_M = \lambda - (\lambda - \mu)(M_t/M_\infty)$. The rationale is that new copies, being very expensive, were produced individually to order, as librarians' book lists and correspondence suggest (*4*), not by the batch on speculation, like printed books. Suppose also that the actual per capita death rate μ is constant, the rationale being that manuscripts' destruction generally was unintentional and hence tended to proceed at a more or less constant per capita rate, barring proscription or obsolescence of the text.

What emerges is the Verhulst-Pearl logistic equation for population growth, which Price suggested might apply to growth of the scientific literature (*5*), and a somewhat different expression for the surviving number of manuscripts m_t (as opposed to the generally unknown M_t): $m_t/m_T = [e^{-\lambda(T-\tau)} + e^{-\mu(T-\tau)}]/[e^{-\lambda(t-\tau)} + e^{-\mu(t-\tau)}]$, where τ is the expected time for a population to grow from $M_0 = 1$ to $M_\tau = M_\infty/2$, and $T = t_T - t_0$ is a reference time interval (*1*) (Fig. 1A). The "fossil" population's age distribution (m_t) should be very sensitive to the probability ratio μ/λ (*1*). If $\mu/\lambda < 3 - 8^{1/2} \approx 0.17$, the cumulative age distribution (m_t) will have an inflection point, and the corresponding age-frequency curve (dm_t/dt) will preserve a distinct maximum corresponding to the maximum in the growth-rate curve (dM_t/dt). Otherwise, if $\mu/\lambda > 3 - 8^{1/2}$, both curves will be concave upward.

The Venerable Bede's *De Temporum Ratione* (725 A.D.) (*6*–*8*) is nearly ideal for studying the dissemination of technical information: It was the standard textbook on arithmetical and calendrical calculation from the eighth to sixteenth centuries (*9*, *10*): Its first chapter, on finger counting, was the standard instruction manual for the "pocket calculator" of the time, and so found a readership in commercial as well as learned circles (*11*); it was ideologically neutral; it is relatively common; and its manuscript tradition is well known (Fig. 1B). Much the same can be said of Bede's three other technical works (*1*), which are included for comparison (Fig. 1, B and C).

Despite many potential complications, the model evidently applies well to all four cases, the iteratively fitted least-squares curves in Fig. 1, B and C, accounting for 96 to 99% of the variance in m_t (R^2, Table 1) (*1*). Figures indicate that a title's manuscripts were about 15 to 30 times more likely to be copied as to be destroyed and had a half-life of four to nine centuries, and that population's doubling time was on the order of two to three decades (Table 1). Modest escalations (*5*) in M_∞ took place during the ninth-century Carolingian renaissance [when Bede's textbooks began to be mass produced as part of the program to promote and standardize education and administration throughout the nascent Holy Roman Empire (*11*, *12*)], during the 12th-century renaissance, and again during the Renaissance of the 15th century (Fig. 1, B and C)—correspondingly modest violations of the model's assumption of constant M_∞, but also evidence that manuscripts' age distribution reflects an ongoing balance between supply and demand, as assumed.

For manuscripts in general, the fraction that escaped destruction during social upheavals such as the Viking wars or French Revolution, or through simple neglect after printed editions appeared, has attracted widely varying guesses but hardly any reasoned estimates. The one most generally accepted for titles of Bede's vintage has been that of Bischoff, who suggested that about one in seven manuscripts have survived from ninth-century Carolingian workshops (*13*, *14*). If a roughly six-century half-life is typical of the Bede manuscripts, as Table 1 suggests, that number would be closer to two in seven—remarkably good agreement, considering the uncertainties (*1*).

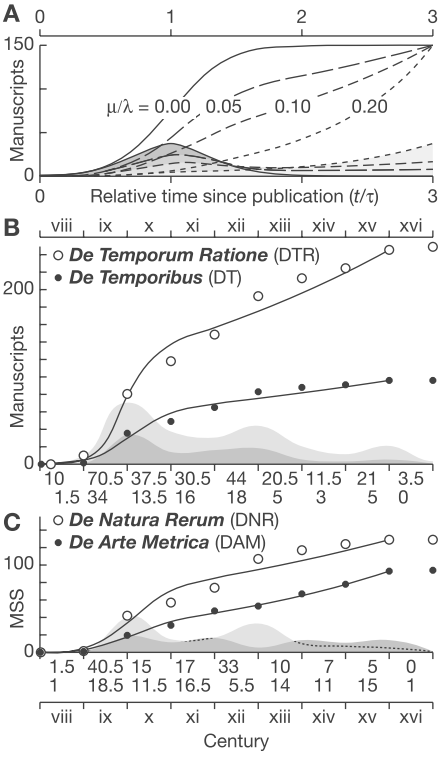


Fig. 1. Age distributions for manuscript populations. (**A**) Age distributions predicted from the model developed for populations with differing "birth" (λ) and "death" (μ) probabilities. Shown are cumulative curves for the number of manuscripts m_t surviving from time t out of a total $m_T = 150$ at a time $t_T = 3\tau$, and each of these curve's time derivatives dm_t/dt, which correspond to the manuscripts' age-frequency curves (*1*). The graph is scaled for comparison with the four cases in (B) and (C). (**B** and **C**) Age distributions for parchment manuscripts of the Venerable Bede's four technical works showing the numbers of manuscripts known from each century and the least-squares curves for m_t up to $t_T = 1500$ A.D. fitted according to the model in (A) (Table 1). The manuscripts themselves range from isolated pages to complete copies. Ages are given to the nearest century because that is how precisely most manuscripts have been dated. The endpoint $t_T = 1500$ A.D. represents the end of the last full century during which copies were produced entirely by hand. The number of copies surviving from each century is given below the age axis for each text and is plotted in the background, tone screens distinguishing the texts (*1*). To alleviate the problems of aliasing, a borderline case dated "ca. 800" or "saec. VIII/IX," for instance, is counted as half a manuscript in the eighth century and half in the ninth. MSS, manuscripts.

Table 1. Estimated values and standard errors of population parameters for the Venerable Bede's four technical works (*1*). cent, centuries; DTR, *De Temporum Ratione*; DT, *De Temporibus*; DNR, *De Natura Rerum*, DAM, *De Arte Metrica*.

Text	λ (cent^{-1})	μ (cent^{-1})	μ/λ	τ (cent)	R^2
DTR	3.6 ± 0.2	0.11 ± 0.00	0.032 ± 0.002	~1.6	0.98
DT	2.4 ± 0.1	0.08 ± 0.01	0.033 ± 0.004	~2.0	0.98
DNR	2.5 ± 0.2	0.10 ± 0.01	0.040 ± 0.006	~2.0	0.96
DAM	2.6 ± 0.0	0.18 ± 0.00	0.070 ± 0.006	~1.9	0.99

The fraction of titles that have survived from Antiquity has been even more uncertain. Present results suggest that a substantial proportion of the more popular texts circulating in the early Middle Ages probably has survived. If a dozen parchment copies of any one of the four Bede texts existed in Carolingian times, and if each copy has had a one-in-seven chance of surviving to the present, the likelihood is greater than 80% that at least one of these would survive today, and thus that the text itself would survive. If a population of a potentially popular text such as those in Table 1 started with $M_0 = 1$, so that the unlimited birth-and-death process would apply as an approximation, the text would have roughly a $\mu/\lambda \leq 0.07$ chance of ultimate extinction (*3*). Thus, if all 35 of the mostly nontechnical works that Bede lists in his *Historia Ecclesiastica Gentis Anglorum* (731 A.D.) (*15*) likewise had an extinction probability of $\mu/\lambda \leq 0.07$, one might expect no more than two or three to be extinct. In fact, three are extinct (*16*), suggesting that present results apply as an approximation to Bede's nontechnical as well as technical works, and quite possibly to contemporary textbooks in general.

Taken together, present results suggest that many if not most scientific and technical works that circulated in Latin on parchment in the early Middle Ages or even in late Antiquity survive in some form. But why then have so few actually survived from Antiquity? For instance, only one of seven works by Pliny the Elder and only a small fraction of the approximately 2000 works on which he based his *Naturalis Historia* (77 A.D.) (*17*), the foremost scientific encyclopedia of the Middle Ages, have survived. The answer may lie in copyists' changeover from papyrus to parchment during the third and fourth centuries (*18*). Surviving texts may be mostly those similar to Pliny's (*19*), which happened to have been in demand during and soon after the transition to the new and more durable medium.

Only further research will tell how accurate or how representative present estimates are. The important point is the apparent feasibility of quantitatively investigating the dynamics of knowledge transmission in ancient and medieval times by demographically analyzing centuries' worth of accumulated paleographic data.

References and Notes

1. Materials and methods are available as supporting material on *Science* Online.
2. J. L. Cisne, in *Encyclopedia of Paleontology*, R. W. Fairbridge, D. Jablonski, Eds. (Dowden, Hutchinson & Ross, Stroudsburg, PA, 1979), pp. 628–635.
3. W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, ed. 2, 1957), vol. 1, pp. 407–411.
4. J. W. Thompson, *The Medieval Library* (Univ. of Chicago Press, Chicago, 1939), pp. 613–629.
5. D. J. de S. Price, *Little Science, Big Science* (Columbia Univ. Press, New York, 1963), chap. 1.
6. C. W. Jones, *Bedae Opera de Temporibus* (Medieval Academy of America, Cambridge, MA, 1943).
7. F. Wallis, *Bede: The Reckoning of Time* (Liverpool Univ. Press, Liverpool, UK, 1999).
8. Bede, *De Temporum Ratione Liber*, C. W. Jones, Ed. (Corpus Christianorum, Series Latina, Brepols, Turnhout, Belgium, 1977), vol. CXXIII B, pp. 237–544.
9. G. Sarton, *Introduction to the History of Science* (Carnegie Institution, Washington, DC, 1927), vol. 1, pp. 503–511.
10. A. Borst, *The Ordering of Time: From the Ancient Computus to the Modern Computer* (Univ. of Chicago Press, Chicago, 1993), chaps. 5 and 6.
11. K. Menninger, *Number Words and Number Symbols*, P. Broneer, Transl. (MIT Press, Cambridge, MA, 1969), pp. 201–208.
12. G. Brown, in *Carolingian Culture: Emulation and Innovation*, R. McKitterick, Ed. (Cambridge Univ. Press, Cambridge, 1994), pp. 1–51.
13. R. McKitterick, *The Carolingians and the Written Word* (Cambridge Univ. Press, Cambridge, 1989), pp. 135–164.
14. J. J. Contreni, in *The New Cambridge Medieval History*, R. McKitterick, Ed. (Cambridge Univ. Press, Cambridge, 1995), vol. 2, pp. 709–757.
15. Bede, *Bede's Ecclesiastical History of the English People*, B. Colgrave, R. A. B. Mynors, Eds., Transl. (Oxford Univ. Press, Oxford, 1969), pp. 566–571.
16. M. L. W. Laistner, H. H. King, *A Hand-list of Bede Manuscripts* (Cornell Univ. Press, Ithaca, NY, 1943), pp. 154–159.
17. Pliny, *Natural History*, H. Rackham, Transl. (Harvard Univ. Press, Cambridge, MA, 1991), preface, pp. 2–21.
18. B. Bischoff, *Latin Palaeography*, D. Ó Crónín, D. Ganz, Transl. (Cambridge Univ. Press, Cambridge, 1990), pp. 7–11.
19. L. D. Reynolds, in *Texts and Transmission*, L. D. Reynolds, Ed. (Oxford Univ. Press, Oxford, 1983), pp. 307–316.

20. I thank R. Ziomkowski, F. Vermeylen, P. Stevens, C. Labandeira, F. King, B. Fisher, J. Chiment, and two anonymous reviewers for help and encouragement. Any errors are mine.

Supporting Online Material

www.sciencemag.org/cgi/content/full/307/5713/1305/DC1
Materials and Methods
SOM Text
Figs. S1 and S2
Table S1
References and Notes

31 August 2004; accepted 30 November 2004
10.1126/science.1104718

The Restoration Potential of the Mesopotamian Marshes of Iraq

Curtis J. Richardson,[1]* Peter Reiss,[2] Najah A. Hussain,[3] Azzam J. Alwash,[4] Douglas J. Pool[5]

Uncontrolled releases of Tigris and Euphrates River waters after the 2003 war have partially restored some former marsh areas in southern Iraq, but restoration is failing in others because of high soil and water salinities. Nearly 20% of the original 15,000-square-kilometer marsh area was reflooded by March 2004, but the extent of marsh restoration is unknown. High-quality water, nonsaline soils, and the densest native vegetation were found in the only remaining natural marsh, the Al-Hawizeh, located on the Iranian border. Although substantially reduced in area and under current threat of an Iranian dike, it has the potential to be a native repopulation center for the region. Rapid reestablishment, high productivity, and reproduction of native flora and fauna in reflooded former marsh areas indicate a high probability for successful restoration, provided the restored wetlands are hydraulically designed to allow sufficient flow of noncontaminated water and flushing of salts through the ecosystem.

The Mesopotamian marshes of southern Iraq (30° to 33°N, 45° to 48°E) are considered by many to be the "cradle of western civilization" and are often referred to as the Garden of Eden (*1*, *2*). Their ecological and cultural value derive from their large expanse of wetland habitat in southwest Asia, once covering nearly 15,000 km^2 (*3*, *4*) (Fig. 1A). Unique to the Mesopotamian marshes is the indigenous population of marsh dwellers, commonly called the Marsh Arabs, who already have a special place in the anthropological and cultural literature for their alluring way of life, living in harmony in relative isolation within the marsh environment for the past 5000 years on man-made reed islands and along the periphery of the marshes (*1*, *5*, *6*). Notably, the marshes were once the permanent habitat for millions of birds and a flyway for millions more migrating between Siberia and Africa (*7*, *8*).

[1]Duke University Wetland Center, Nicholas School of Environment and Earth Sciences, Duke University, Box 90333, Durham, NC 27708, USA. [2]Development Alternatives, 7250 Woodmont Avenue, Bethesda, MD 20895, USA. [3]College of Science, University of Basra, Basra, Iraq-Basrah-Garmet Ali-P.O. Box 49, Iraq. [4]Iraq Foundation, 1012 14th Street N.W., Washington, DC 20005, USA. [5]International Resources Group, 1211 Connecticut Avenue N.W., Washington, DC 20036, USA.

*To whom correspondence should be addressed. E-mail: curtr@duke.edu

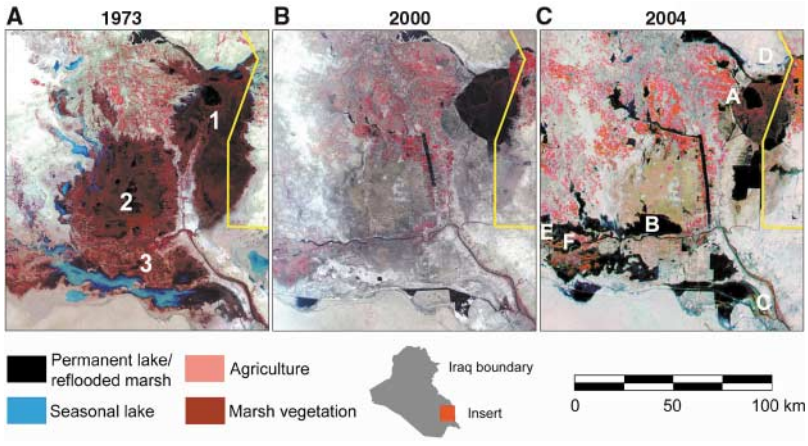


Fig. 1. (**A**) A composite satellite view of the Mesopotamian marshlands from Earth observation Landsat scenes taken in 1973 to 1976 from a mosaic of 4 Landsat 1 and 2 false-color, near-infrared images (*3*). Dense marsh vegetation [mainly *P. australis* based on field observations in 1970s (*29*)] appears in dark red, seasonal lakes in blue, agriculture in pink, and permanent lakes in black. The red elongated patches along riverbanks are date palms. The three main marsh areas are the Al-Hawizeh, Central marsh, and Al-Hammar, labeled on the map as 1, 2, and 3, respectively. (**B**) In this Landsat 7 Enhanced Thematic Mapper (*3*) mosaic taken in 2000, most of the drained marshes appear as grayish brown patches indicating dead marsh vegetation or low shrub desert plants and dry ground. The very light to gray patches are bare areas with no vegetation and in some areas indicate salt evaporites or shells covering the bottoms of former lakes as we discovered in 2003. The drained former marsh area (shown in brown, gray, and white) indicates that 85% of the 8926 km^2 of permanent marsh shown in 1973 marshlands (A) had been destroyed by 2000. Only 3% of the Central marsh and 14.5% of the Al-Hammar remained in 2000 (*3*). The largest expanse (≅1025 km^2) of the remaining natural marsh Al-Hawizeh near the Iranian border is shown in dark red. To complete this destruction the Glory River, shown as a straight line canal across the top and down the east side of the Central marsh, was constructed by the Hussein Regime in 1993 to completely dry up the Central marsh by stopping water inflow from the Tigris River. (**C**) February 2004 false-color image of the remaining Mesopotamian marshlands show the areas newly reflooded since the war in black (courtesy of Moderate-Resolution Imaging Spectroradiometer Rapid Response Project at NASA). Reflooded areas adjacent to the Al-Hawizeh, the western area of the Al-Hammar, and waterways in the northern and southern parts of the Central marshes are also clearly visible in black. The Al-Hawizeh (called Hawr Al-Azim by Iran) is the best remaining natural marsh in the region (site A). It straddles the Iraq-Iran border (yellow line). During a field survey in February of 2004, we discovered an Iranian dike under construction that, if completed, will traverse directly through the Al-Hawizeh marsh, along the Iraq/Iranian border and, as a result, will notably reduce the water input from the Karkheh and Karun rivers to the marsh. The ecological affects of this massive water diversion are unknown, but it will substantially impact the last remaining natural marsh system in Iraq. Our sampling sites A to F are indicated on (C) with matching ground photos shown in Fig. 2, A and B, and fig. S1, A to D.

Today, less than 10% of the marshes in Iraq remain as fully functioning wetlands because of extensive drainage and upstream agricultural irrigation programs on the Tigris and Euphrates rivers implemented during Saddam Hussein's regime, particularly from 1985 to 2000 (*3*, *7*). Remnant marshes are located mainly in the northern portion of Al-Hawizeh, which straddles the Iran-Iraq border (Fig. 1A, site 1). Two other marshes, the Central and Al-Hammar located to the west and south, respectively (Fig. 1A, sites 2 and 3), were almost totally desiccated by 2000 (Fig. 1B) (*9*). This environmental disaster has been compared in scale to the drying up of the Aral Sea in Central Asia and to the deforestation of the Amazon (*3*). If substantial portions of the marshes are not restored quickly, 66 bird species may now be at risk (*3*, *8*). Coastal fisheries in the Persian Gulf, which used the marshlands for spawning migrations and nursery grounds for penaeid shrimp (*Metapenaeus affinis*), have drastically declined in numbers (*9*). The marshlands also once served as a natural filter for waste and other pollutants in the Tigris and Euphrates rivers, protecting the Persian Gulf, which is now noticeably degraded along the coast of Kuwait (*3*, *7*, *10*).

Because of limited access until 2003, little is known about the current water quality of river discharges, the ecological conditions of remaining marsh areas, or the potential for the restoration of the drained marshes. This problem was compounded by a loss of historic hydrologic flow data during the looting of local offices of the Ministry of Water Resources. It has been postulated that only 15 to 20% of the drained marshes can be restored as a result of excessive salt buildup, pollution, the dumping of toxic wastes and poisonings during the war, a severe reduction in available water, and a loss of the seed bank of native plant species (*3*, *11*, *12*). Moreover, the long-term quantity of water and sediment available to restore the marshes is uncertain because construction of 32 dams on the Tigris and Euphrates, mainly since the 1960s, substantially altered the marsh hydrology even before Hussein's massive drainage program of the 1990s (*3*). Complicating the marsh restoration are projects such as the massive Ataturk Dam built in Turkey in 1998, which has a reservoir capacity greater than the Euphrates annual flow of 30.7 billion cubic meters at the Turkish border (*3*). A massive dike project observed in the field in February of 2004 along the Iran-Iraq border will add to the problem by cutting off the main water supply to the only natural remaining marsh, the Al-Hawizeh.

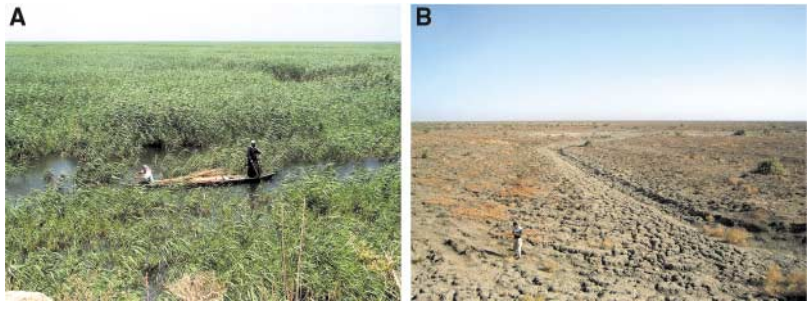


Fig. 2. (**A**) Marsh Arab fishermen collecting reeds (*P. australis*) in the natural Al-Hawizeh marsh (lat 31°38′583″N, long 47°35′203″E) near the Iranian border in June of 2003. (**B**) A view of the totally drained Central marsh near Chibayish (lat 30°58′102″N, long 47°09′033″E) in June of 2003. An Iraqi engineer from the Ministry of Water Resources is viewing the cracked and desiccated marsh soil adjacent to a dried out streambed.

Here, we present an analysis, beginning in June 2003, of soils, water, and ecological conditions for the remaining natural, reflooded, and the totally drained marshes along with an assessment of water constituents in the Tigris and Euphrates upstream and the Shatt Al-Arab downstream of the marshes. The effects of extensive drainage and uncontrolled reflooding on water quality, soil conditions, and ecological recovery are the focus of this study.

Major differences occur in the physical and chemical characteristics (*13*) of the natural (Fig. 2A), drained (Fig. 2B), and reflooded (fig. S1, A and B) marsh soils, as shown by principal component analysis (PCA) in Fig.

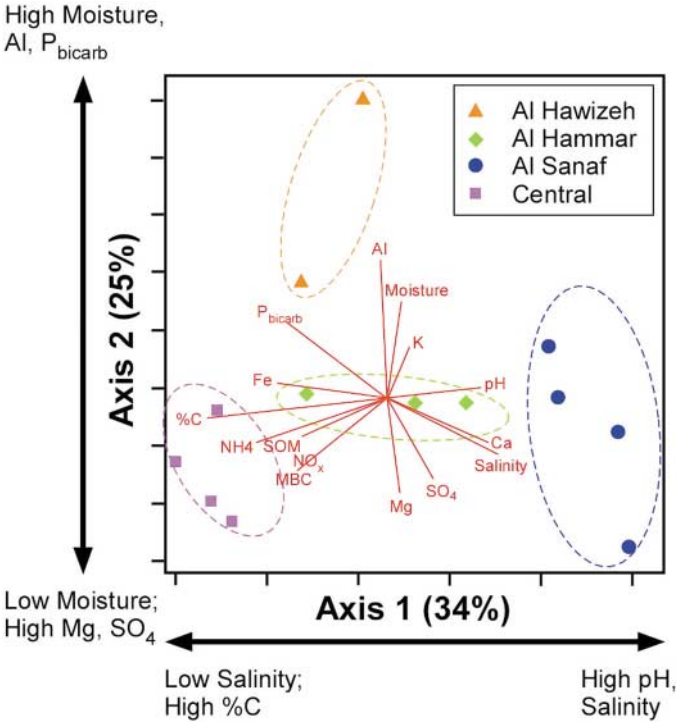


Fig. 3. A PCA of the marsh soils of southern Iraq. Axis 1 accounted for 33.9%, axis 2 accounted for 25.4%, and axis 3 accounted for 13.0% of the total variance. A number of soil properties, including salinity (0.31) and exchangeable calcium (Ca) (0.31), had strong positive loadings on axis 1 and showed distinct differences among marsh sites. Soil properties with strong negative loadings on axis 1 included percentage C (–0.40) and soil organic matter (–0.33) as well as extractable Fe (Fe) (–0.32). Soil moisture (0.33) had a strong positive loading on axis 2. Strong negative loadings on axis 2 occurred for soil organic matter (SOM) (–0.43), MBC (–0.42), NH_4-N (–0.35), and No_x-N (–0.35).

3. PCA on axis 1 separated the reflooded Al-Sanaf marsh (fig. S1B), a diked wetland with no outflow and dominated by stands of the halophyte glasswort (*Salicornia* spp.), from the totally drained Central marsh (Fig. 2B), which was barren of wetland species, primarily on higher soil salinities (26.0 ± 18.6 dS/m versus 10.6 ± 0.9 dS/m, where errors are the standard deviation), lower percentage of carbon (3.5% ± 0.2 versus 7.6% ± 0.6), and soil organic matter (SOM) (7.5 ± 2.2% versus 15.1 ± 1.1%) (Fig. 3). The higher SOM content in the Central marshes (Fig. 2B) supports earlier reports (*1*, *14*) that this area with deeper lakes historically had high plant productivity, with reeds growing up to 8 m in height. Surprisingly, the soil organic matter has not been completely oxidized after nearly a decade of drainage. The highest microbial biomass carbon (MBC) (842 ± 185 mg C/kg soil) was also found in the drained Central marsh, whereas the MBC values of the natural and reflooded marshes were 50% lower. The Central marsh had the highest total soil N and second highest P content (0.53 ± 0.04% N, 794 ± 86 μg/g P) as compared to the Al-Hawizeh (0.30 ± 0.10% N, 935 ± 388 μg/g P), the Al-Hammar (0.27 ± 0.15% N, 618 ± 104 μg/g P), and the Al-Sanaf (0.06 ± 0.03% N, 507 ± 58 μg/g P). We encountered the most vigorous and vast monocultures of native reed (*Phragmites australis*) communities in the Al-Hawizeh (Fig. 2A). This marsh had the highest levels of readily available bicarbonate extractable phosphorus (130 ± 57 μg/g) and soil moisture (41.7 ± 0.3%) and the lowest salinity (1.8 dS/m). The natural marsh was separated on PCA axis 2 from the disturbed marshes along a gradient correlated with Al, Mg, P sulfate (SO_4), and moisture (Fig. 3). These baseline soil chemistry differences support a different geological parent material for the Al-Hawizeh, as suggested by Buringh (*15*). The eastern Al-Hammar marsh (fig. S1A) has high residual soil salinities (17.5 ± 18.8 dS/m) and high sulfur content (extractable SO_4-S, 1.0 ± 0.9 mg/g), which may be responsible for the lack of reestablishment of native vegetation in reflooded areas as of March 2004.

As a second step, we analyzed for the presence of pesticides, polychlorinated biphenyls, polycyclic aromatic hydrocarbons, and potentially toxic metals and assessed salinity of the soils in all four marshes (*13*) to evaluate constraints on future marsh restoration. We found no detectable concentrations of any organic xenobiotics in the marsh soils, which was similar to a study on downstream sediments in 1997 to 1998 (*10*) but in contrast to reports of serious pesticide pollution in the 1980s (*12*). Metal concentrations (*13*) were within normal ranges reported earlier for the marsh soils of Iraq (*14*, *15*). Soil salinity, however, exceeded 16 dS/m in both the Al-Sanaf and eastern portion of Al-Hammar marshes. These levels are severe and suitable only for plants highly tolerant of saline or sodic conditions (*16*, *17*). The presence of main species like halophytic saltbush (*Atriplex* spp.), Athel tree (*Tamarix aphylla*), and glasswort in these two former marshes (fig. S1, A and B) also suggests a soil salinity problem and indicates that these areas will not be easily restored to native marsh vegetation unless saline conditions can be reduced. Only the natural Al-Hawizeh marsh displayed healthy reed vegetation (Fig. 2A) and had salinity concentrations consistently below 2 dS/m, which can be classified as nonsaline. The Central marsh has a wide range (6 to 17 dS/m) of salinities. Salinities ranging from 5 to 10 dS/m will stimulate *Phragmites* seed germination but salinities ≥25 dS/m will inhibit germination (*18*), as well as plant productivity. The yield of most crops such as rice and wheat, which are traditionally grown at the edge of the marshes by the Marsh Arabs (*19*), could also be severely reduced at salinities >8 dS/m (*20*, *21*).

We sampled surface water quality at a number of upstream locations on the Tigris and Euphrates rivers to examine whether concentrations of inorganic and organic chemicals would be detrimental to marsh restoration as well as to determine compatibility with marsh and downstream waters in the Shatt Al-Arab (table S1) (*13*, *22*) [supporting online material (SOM) text]. All upstream and marsh surface waters were highly oxygenated, but O_2 was significantly ($P < 0.05$) reduced in the Shatt Al-Arab, where untreated wastewater is being released from cities, including Basra. Salinity, conductivity, and total dissolved solids (TDS) values were low, and pH was above 7.5 at all sites except for the enclosed Al-Sanaf, where these variables were significantly higher ($P < 0.05$). These findings suggest that the restricted water outflows in the Al-Sanaf coupled with regional evapotranspiration rates in excess of 245 cm/year (*15*) have resulted in extremely high ion concentrations, pH, and TDS values (table S1), similar to those found on highly salinized portions of the Jordan River (*23*). Similar salinity trends in the soils data (Fig. 3) provide strong evidence that even the release of high-quality waters into former marshlands with existing saline soils and without adequate flow will result in failed marsh restoration because of the creation of a highly toxic saline-sodic environment.

Concentrations of most minor constituents (such as Cu, Fe, Ni, and Al) and trace metals (such as As, Cd, Pb, and Rb) (*13*) were low except in the enclosed Al-Sanaf. Of major concern was our finding of high concentrations of Se (12.3 ± 6.4 μg/liter) in Al-Sanaf and in upstream waters, which were more than twice the recommended U.S. national water quality criterion of 5 μg/liter (*24*, *25*). The source of Se is agricultural soil waters drainage, which is then concentrated by evapotranspiration in confined marshes. Water quality standards for Se in both fresh and marine water have been set at 2 μg/liter in many parts of the world to prevent

reproductive failures in fish and birds caused by higher food-chain bioaccumulation (*26*). Our findings of high Se in Al-Sanaf and other marsh waters suggest that the placement of waters into fully diked areas may set in motion an insidious mode of toxicity that could have serious consequences for marsh restoration, migratory bird populations, and future agriculture unless prudent management actions are taken in the years ahead to avoid the total containment of waters into holding areas, which concentrate ions under the extremely high evaporative desert conditions.

In February 2004, we established monitoring stations (*13*) in the reflooded marshes of Abu Zarag (western Central marsh, site E in Fig. 1C, and fig. S1C) and Suq Al-Shuyukh (western part of Al-Hammar, site F in Fig. 1C, and fig. S1D) to assess the ecological recovery and restoration status of two former wetlands reflooded in 2003. With Iraqi scientists, we analyzed water quality and surveyed the dominant flora and fauna in the natural Al-Hawizeh to compare with the two reflooded marshes (*13*). Salinity and conductivity values closely followed the dominant cation concentrations ($Na^{+1} > Ca^{+2} > Mg^{+2} > K^{+1}$), which were significantly lower ($P < 0.05$) in Abu Zarag than in Al-Hawizeh (except for Ca^{+2} and Mg^{+2}) and Suq Al-Shuyukh (table S2). Suq Al-Shuyukh had the highest concentrations of most constituents, which indicates that this reflooded site is more saline and chemically enriched compared with the other two sites (table S2). Notably, Se concentrations were extremely low in both Abu Zarag and Al-Hawizeh and were within recommended standards (*24*–*26*) at Suq Al-Shuyukh. Overall water quality at Abu Zarag and Al-Hawizeh was similar because both now receive higher quality Tigris waters. However, our current water chemistry values (table S2), when compared with a historical survey completed before drainage in the Al-Hammar marsh (*14*, *27*), revealed an increase in conductivity (240%), TDS (140%), Na^{+1} (170%), Mg^{+2} (158%), Ca^{+2} (240%), Cl^{-1} (160%), and HCO_3^- (180%) in the Suq Al-Shuyukh region during the past 20 years. In contrast, measured salinities in 1981 (*14*) from seven locations in the Central marsh averaged 0.6 (±0.4) mS/cm, values very similar to our current measurements at Abu Zarag (table S2). Collectively, our studies indicate that the reflooded portions of the western Al-Hammar at Suq Al-Shuyukh (table S2) and eastern Al-Hammar (Fig. 3) have experienced substantial increases in saline conditions. The long-term effects of this level of salinity increase are unknown, but present levels are within the normal variation between wet season and dry season found with the marshes (*28*, *29*). The cause for this marked increase is unknown but probably relates to a rise in salinity in the Euphrates as well as increased flux into the water column of ions concentrated in the soil after 10 years of drainage and evaporation. In the case of eastern Al-Hammar, an increased influence of tidal seawater flowing from the Shatt Al-Arab into the marsh has also occurred from breaches in dikes in 2003 and may be preventing reestablishment of freshwater marsh species (fig. S1A).

Our summer ecological survey (*13*) (SOM text) completed one year after reflooding indicates that water flowing into the marshes from the Euphrates and especially the Tigris is of higher quality than we originally hypothesized, and as a result, early successional stages of marsh restoration are occurring in a number of drained former marshes (Fig. 1C, sites E and F, and fig. S1, C and D). The reflooding of Abu Zarag with Tigris water has resulted in better water quality, higher algal productivity, and more diversity of plant and bird species than at Suq Al-Shuyukh, which receives higher salinity Euphrates water (tables S2 and S3). Both marshes show partial reestablishment of many of the dominant plant species, but biomass and species numbers are low compared with historical records at Suq Al-Shuyukh (*29*). Chlorophyll a concentrations reveal mesotrophic conditions (table S3) at all sites, which are within historic values except for Suq Al-Shuyukh (*14*). Abu Zarag, the site with the lowest salinity, Na^{+1}, Cl^{-1}, and TDS has the highest chlorophyll a concentrations and concomitantly the highest O_2 production (tables S2 and S3). Our macrophyte analysis indicates that *P. australis*, as in the past, is the dominant macrophyte species. Many key amphibian, fish, bird, and mammal species are not yet present or, if present, are in lower numbers, especially at Suq Al-Shuyukh (table S3).

We conclude that restoration is proceeding in the two marshes but at different rates and species composition. Unknown is the fate of the many highly threatened species (*2*, *3*) of the marshlands and the future of these restored marshes as a major flyway for Asia's wintering wildfowl. Here, the quality of water and food supplies as well as the diversity of communities, including mature habitats, will be critical to the return of many of these species. For example, our main observations of otter (*Lutra lutra*) were only from the natural Al-Hawizeh. The high water quality, low salinity, and the presence of permanent lakes and dense vegetation in the remaining natural Al-Hawizeh marsh give hope that this area can function both as a refugium and native repopulation center for the region, provided it is not desiccated by Iran's dike construction activities.

Because future water supplies for marsh restoration may be limited by upstream dams, soil conditions need to be carefully monitored before reflooding new areas to prevent the establishment of saline-sodic soil conditions that will impede marsh revegetation, as seen at eastern Al-Hammar (fig. S1A) and Al-Sanaf (fig. S1B). Severe water quality degradation may be avoided in some reflooded marsh areas if they are hydraulically designed to allow enough high-quality water to flush salts from the ecosystem. Of concern from our study is the potential for the bioaccumulation of Se up the food chain, which could result in severe toxic effects in the marshes for higher trophic levels (*24*). Finally, it is unknown if sufficient water supplies can be made available, especially in drought years, to maintain long-term successful marsh restoration over large areas. However, the high quality of water, the existing soil conditions, and the presence of stocks of native species in some regions indicate that the restoration potential for a substantial portion of the Mesopotamian marshes is high. The stakes are also high: The future of the 5000-year-old Marsh Arab culture and the economic stability of large portions of southern Iraq are dependent on the success of this restoration effort.

References and Notes

1. W. Thesiger, *The Marsh Arabs* (Longman, London, 1964).
2. E. Nicholson, P. Clark, Eds., *The Iraqi Marshlands: A Human and Environmental Study* (Politico's Publishing, London, ed. 2, 2003).
3. H. Partow, "Demise of an ecosystem: The disappearance of the Mesopotamian marshlands" [United Nations Environment Programme (UNEP) Publication UNEP/DEWA/TR.01-3, Nairobi, Kenya, 2001].
4. J. Brasington, in *The Iraqi Marshlands: A Human and Environmental Study*, E. Nicholson, P. Clark, Eds. (Politico's Publishing, London, 2002), pp. 151–168.
5. S. M. Salim, *Marsh Dwellers of the Euphrates Delta* (Univ. of London Press, London, 1962).
6. G. Young, *Return to the Marshes: Life with the Marsh Arabs of Iraq* (Collins, London, 1977).
7. E. Maltby, Ed., "An environmental and ecological study of the marshlands of Mesopotamia" (Draft Consultive Bulletin, Amar Appeals Trust, London, 1994).
8. M. I. Evans, in *The Iraqi Marshlands: A Human and Environmental Study*, E. Nicholson, P. Clark, Eds. (Politico's Publishing, London, 2002), pp. 201–219.
9. *Environment in Iraq: UNEP Progress Report* (UNEP, Geneva, 2003).
10. T. Saeed, A. N. Al-Ghandban, H. Al-Shemmari, M. Al-Mutairi, H. Al-Hashash, *Water Sci. Technol.* **40**, 89 (1999).
11. A. DouAbul, H. al-Saad, A. Al-Timari, H. Al-Rekabi, *Arch. Environ. Contam. Toxicol.* **17**, 405 (1988).
12. "Eden again international technical advisory panel report" (Iraq Foundation, Washington, DC, 2003).
13. Materials and methods are available as supporting material at *Science* Online.
14. B. K. Maulood, G. C. F. Hinton, B. A. Whitton, H. A. Al-Saadi, *Hydrobiology* **80**, 269 (1981).
15. P. Buringh, *Soils and Soil Conditions of Iraq* (Ministry of Agriculture–Directorate General of Agricultural Research and Projects, Baghdad, 1960).
16. M. A. El Demerdash, A. K. Hegazy, A. M. Zilay, *Vegetatio* **112**, 141 (1994).
17. M. S. Serag, A. H. A. Khedr, *Environmetrics* **12**, 219 (2001).

18. L. A. Meyerson, K. Saltonstall, L. Windham, E. Kiviat, S. Findlay, *Wetl. Ecol. Management* **8**, 89 (2000).
19. A. H. Al-Mousawi, B. A. Whitton, *Arab Gulf J. Sci. Res.* **1**, 237 (1983).
20. L. A. Richards, Ed., "Diagnosis and improvement of saline and alkali soils" (Agriculture Handbook No. 60. U.S. Department of Agriculture, Washington, DC, 1954).
21. E. V. Maas, *Appl. Agric. Res.* **1**, 12 (1986).
22. E. V. Maas, S. R. Grattan, in *Agricultural Drainage*, R. W. Skaggs, J. Van Schilfgaarde, Eds. (American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI, 1999), pp. 55–110.
23. E. Farber *et al.*, *Geochim. Cosmochim. Acta* **68**, 1989 (2004).
24. A. D. Lemly, *Hum. Ecol. Risk Assess.* **5**, 1139 (1999).
25. A. D. Lemly, *Environ. Monit. Assess.* **28**, 83 (1993).
26. A. D. Lemly, *Ecotoxicol. Environ. Saf.* **10**, 314 (1985).
27. K. M. Banat, F. M. Howari, *J. Arid Environ.*, in press.
28. N. A. Hussain, Ed., *Ahwar of Iraq: An Environmental Approach* (Marine Science Center, Basra University, Basra, Iraq, 1992).
29. M. R. Al-Hilli, thesis, University of Cairo (1977).
30. Funding for this project was provided by the U.S. Agency for International Development through the Iraq Marshlands Restoration Program, implemented by Development Alternatives, Bethesda, MD. The Iraqi scientists at the University of Basra and Baghdad who contributed data for the development of table S3 are regional experts in algae, aquatic plants, fisheries, macroinvertebrates, and mollusca. Given present security concerns about their safety, we have not included their names here. Thanks to A. Douabul of the Iraq Foundation for providing data on Abu Zarag. W. H. Schlesinger, R. S. King, R. DiGiulio and J. W. Pahl reviewed the manuscript at its final stages. Finally, thanks to M. Ho, G. L. Bruland, and R. L. Neighbarger for statistical, graphics, and technical assistance and to H. Partow for helpful discussions and Landsat data on the areas of marsh reflooding.

Supporting Online Material
www.sciencemag.org/cgi/content/full/307/5713/1307/DC1
Materials and Methods
SOM Text
Figs. S1 and S2
Tables S1 to S3
References

28 September 2004; accepted 6 January 2005
10.1126/science.1105750

Genotypic Diversity Within a Natural Coastal Bacterioplankton Population

Janelle R. Thompson,[1,2] Sarah Pacocha,[1,2] Chanathip Pharino,[1] Vanja Klepac-Ceraj,[1,2] Dana E. Hunt,[1] Jennifer Benoit,[1,2] Ramahi Sarma-Rupavtarm,[1] Daniel L. Distel,[3] Martin F. Polz[1]*

The genomic diversity and relative importance of distinct genotypes within natural bacterial populations have remained largely unknown. Here, we analyze the diversity and annual dynamics of a group of coastal bacterioplankton (greater than 99% 16*S* ribosomal RNA identity to *Vibrio splendidus*). We show that this group consists of at least a thousand distinct genotypes, each occurring at extremely low environmental concentrations (on average less than one cell per milliliter). Overall, the genomes show extensive allelic diversity and size variation. Individual genotypes rarely recurred in samples, and allelic distribution did not show spatial or temporal substructure. Ecological considerations suggest that much genotypic and possibly phenotypic variation within natural populations should be considered neutral.

Molecular evidence increasingly demonstrates the remarkable genetic diversity of the microbial world (*1*, *2*), yet ecological interpretation of this diversity remains elusive. This is largely because microbiologists rely on studies of clonal isolates or environmental gene libraries to infer biogeochemical and pathogenic functions of natural bacterial populations. What is missing, however, is quantitative information regarding the environmental prevalence of individual genotypes that would allow inference of their ecological importance or competitive success. It may be expected that ecologically distinct populations display relatively high clonality, because bacterial genomes have a high potential for adaptive mutations, which may lead to purging of genotypic diversity from within the population by selective sweeps (*3*, *4*). However, this view is increasingly difficult to reconcile with recent observations of high levels of differentiation among closely related genomes [e.g., (*5*, *6*)] and the recovery of vast numbers of similar but nonidentical homologous genes from environmental samples (microdiversity) (*7*–*9*). Therefore, questions include whether competition among individual strains is strong enough to result in frequent selective sweeps or instead whether natural populations accumulate large neutral allelic and perhaps even genomic variation (*8*). However, the diversity and prevalence of individual variants within environmental bacterial populations has not been extensively explored, and so questions regarding the ecological importance of genotypic variation remain unanswered.

To analyze genotypic diversity and overall population size quantitatively, we combined culture-dependent and -independent methods to assess the number, extent of variation, and relative frequency of genotypes within a well-defined natural bacterial population. We chose a coastal assemblage of *Vibrio splendidus*, previously identified as a phylogenetically discrete cluster denoted by nearly identical (<1% divergent) 16S rRNA sequences in an analysis of bacterioplankton community structure (*8*). We have proposed that such ribotype clusters represent ecologically differentiated units, i.e., ecotypes or populations (*8*). Thus, we defined the *V. splendidus* cluster as a population of naturally co-occurring genomes that can be tracked quantitatively in the environment and identified in strain collections by their distinct rRNA genes.

Quantification of the *V. splendidus* population over an annual cycle by quantitative polymerase chain reaction (QPCR) (*10*) revealed that it is consistently present as a member of the coastal bacterioplankton community and displays seasonal variation in abundance (Fig. 1A). Concomitant with quantification, we isolated strains from five temporal samples on *Vibrio*-selective media and identified strains by 16S rRNA sequence analysis (Fig. 1B) (*11*). Overall, 20 distinct *Vibrio* (and closely related *Photobacterium*) taxa grew on the media (Fig. 1, B and C), but the majority of isolates (232 of 333) were identified as members of the *V. splendidus* population (red sectors in Fig. 1B). This dominance in all collections, except the cold-water sample (March 2003) (Fig. 1B), roughly parallels the culture-independent quantification by QPCR (Fig. 1A).

Determination of sequence diversity of a universally distributed protein-coding gene (*Hsp60*) among all 333 *Vibrio* isolates showed high heterogeneity but confirmed the monophyly of the *V. splendidus* population detected by the rRNA sequence analysis (*12*). We observed 141 different *Hsp60* alleles among the 232 *V. splendidus* isolates (Fig. 2), and extrapolation using the Chao-1 richness estimator (*13*) suggests a minimum of 436 alleles in

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [2]Department of Biology, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA. [3]Ocean Genome Legacy, New England Biolabs, Beverly, MA 01915, USA.

*To whom correspondence should be addressed. E-mail: mpolz@mit.edu

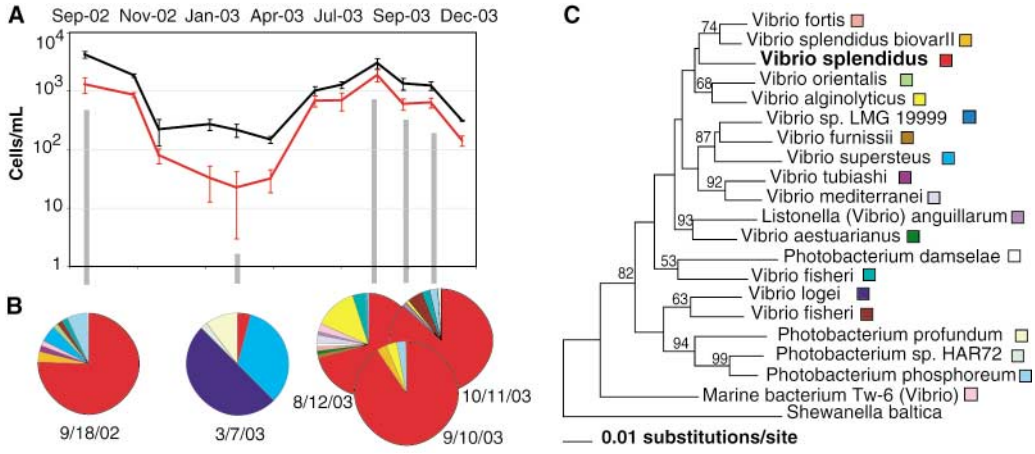


Fig. 1. Diversity and abundance of coastal vibrioplankton (Plum Island Sound, MA) in monthly samples taken over an entire year. (**A**) Quantification of *V. splendidus* (red) and total vibrio (black) populations by culture-independent QPCR. (**B**) Relative proportion of *Vibrio* and *Photobacterium* isolates by phylogenetic association. Color codes correspond to those in (C). (**C**) Phylogenetic relationships among representative *Vibrio* and *Photobacterium* isolates inferred from distance analysis of partial 16S rRNA sequences (bootstrap proportions >50% are indicated above nodes).

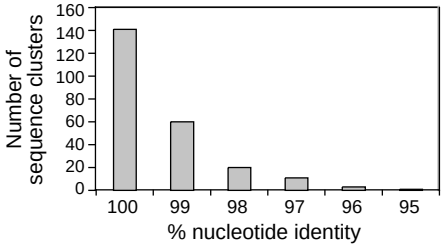


Fig. 2. Number of distinct *Hsp60* clusters among *V. splendidus* isolates observed as cluster cutoff values are decreased from 100 to 95%.

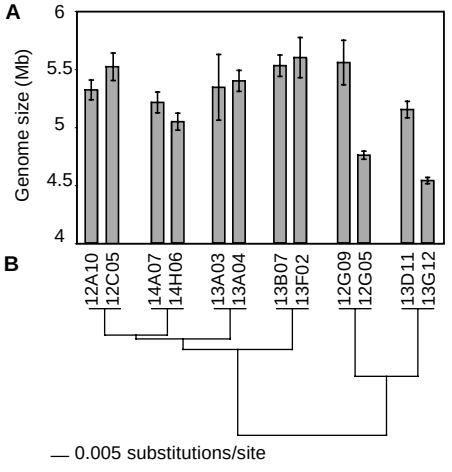


Fig. 3. Genome size estimates and phylogenetic relationships of *Hsp60* sequences for 12 *V. splendidus* isolates chosen as pairs with identical *Hsp60* alleles, encompassing all levels of *Hsp60* variation observed in the strain collection. (**A**) Genome sizes determined by PFGE as averages of six independent estimates, each obtained from single enzyme digests run to resolve large-, medium-, and small-sized bands, respectively, and repeated three times for each of two enzymes (NotI/SfiI or NotI/AscI) per isolate. (**B**) Phylogenetic relationships of *Hsp60* alleles inferred from maximum likelihood analysis with assumption of molecular clock from partial gene sequences. Isolate identifiers correspond to month (12 indicates 12 August 2003; 13, 10 September 2003; 14, 11 October 2003) of isolation and strain name.

the total sampled volume (31.5 ml). Despite these high numbers, the *Hsp60* sequences collapse into a single group at 95% nucleotide consensus (Fig. 2 and fig. S1), with variation primarily limited to neutral third-codon positions [average ratio of nonsynonymous (K_A) to synonymous (K_S) substitutions, K_A/K_S = 0.04]. No single allele showed clear dominance among the isolates, suggesting a relatively even abundance of the strains (fig. S1), and analysis of molecular variance did not reveal evidence for population structure associated with temporal (1 month to 1 year) and spatial (100 µl to 2 ml) scales.

Still greater heterogeneity was revealed within the *V. splendidus* population when the number of distinguishable genotypes among the cultured strains was assayed by pulse field gel electrophoresis (PFGE). PFGE analysis detected 180 different genome patterns among the 206 strains tested, demonstrating that the majority of isolates possess distinct genomes (fig. S1). The Chao-1 estimator (*13*) yielded a total of 1287 genotypes in the samples, considerably exceeding the estimated allelic diversity of the *Hsp60* gene. Moreover, the PFGE analysis provides evidence that, in addition to accumulation of point mutations as observed for the *Hsp60* alleles, a large proportion of genotypes are differentiated by insertions and deletions of large genome fragments. This was suggested by variation among strains in the sum of genome-fragment sizes in the PFGE gels and was explored in detail for a set of 12 strains. Pairs of isolates with identical *Hsp60* alleles were chosen so that sequence identities between pairs reflected overall divergence in these genes (Fig. 3). Among these 12 isolates, genome sizes ranged from 4.5 to 5.6 Mb with only weak correlation of genome size difference to *Hsp60* sequence divergence (r = 0.37) and only minor contribution to size estimates by plasmids (*12*). Even when comparing strains with identical *Hsp60* alleles, most (four of the six pairwise comparisons) showed significant genome size differences (Student's t test, $P < 0.01$; ~170 to 800 kb variation), indicating that specific alleles may be poor markers for distinct genomes.

The high degree of heterogeneity observed among the *V. splendidus* genomes suggests that the average concentration of individual genotypes is small in the sampled environment. To illustrate this, we divided the QPCR-based estimates of population size of *V. splendidus* in samples taken in August, September, and October 2003 (1890, 600, and 640 cells/ml, respectively) (Fig. 1A) by the Chao-1 estimates for the number of *Hsp60* alleles (125, 94, and 279, respectively) and genotypes (465, 553, and 901, respectively) in those same samples. The result suggests that distinct *Hsp60* alleles occurred in the monthly samples at average concentrations of 2 to 15 cells per ml (or at a frequency of 0.3 to 1%), whereas distinct genotypes were present at ~10-fold lower frequency (average concentration for all samples estimated at <1 cell per ml). If the possibility of isolation bias is taken into account, the estimated concentrations (population size/richness) would be even lower, because isolation bias would lead to an underestimation of richness but would not affect population size estimates by QPCR.

What could explain such high diversity of *V. splendidus* genotypes in this environment? The observed pattern suggests that purging of genotypes from within the population (operationally defined as a ribotype cluster) is rare compared with processes introducing variation and that variation persists because it is either favored by selection (e.g., by balancing selection or niche differentiation) or is neutral. Indeed, some proportion of the observed genotypic diversity may reflect the differentiation of (sub)populations that are specialized to particular environmental con-

ditions in the complex life-style of vibrios (including free-living and animal- or particle-associated states). However, ecological considerations suggest that much of the observed genotypic diversity has little adaptive importance in the context of the water column. Given their low estimated concentration (<1 cell/ml), individual genotypes would occupy much less than a trillionth of the volume of a ml of seawater. Because resources are thought to arise in small patches that are unpredictable relative to the location of any given cell (*14*), access by distinct genotypes to conditions allowing rapid growth may be largely stochastic, relegating strong competitive interactions between genotypes to ephemeral microzones. In addition, top-down interactions like predation (*15*) may quickly erase any localized dominance of genotypes. Thus, although individual genotypes may achieve rapid growth in microzones or microcolonies, averaged over the water column their differences do not result in lasting growth advantage (i.e., they are effectively neutral) and so the observed vast genotypic diversity can coexist.

Previous studies have shown that substantial variation in gene content and genome size may occur among closely related genomes (albeit drawn from separate environments) (*6*, *16–18*). Such variation can arise via gene duplication, insertion, and deletion or by horizontal gene transfer (HGT) mediated by phages, plasmid-borne transposons, and integrons (*19*). In fact, HGT is now regarded as a major source of innovation in bacterial evolution (*20–22*), and several cases of environmental differentiation have been linked to specific gene addition or loss (*23–25*). However, it has also been suggested that most acquired sequences do not confer a selective advantage on their host and can be neutral targets for deletion and mutational events, leading to a dynamic genome (*26*, *27*). Indeed, model results indicate that such neutral genome segments are likely to be transient elements represented in only a small fraction of a population (*28*). Our results expand such previous considerations to indicate that large genome modifications, possibly including HGT, are observed with high frequency in genomes that contain identical rRNA and *Hsp60* sequences and coexist within the same natural population.

It will be important to ask whether such extensive genomic variation is a general feature of natural bacterial populations. We have recently observed that two microbial communities are composed of hundreds of microdiverse ribotype clusters (*8*, *9*) and have proposed that these denote bacterial populations that arise by rare selective sweeps followed by effectively neutral diversification (*8*). Consistent with this hypothesis, we show that one such cluster occurs predictably in the bacterioplankton community and contains extensive diversity, much of which may be neutral in the ecological context of the water column. If similar patterns of diversity are common to bacterial communities, caution should be exercised in interpreting the extent to which gene complements or even metabolic traits of individual isolates may reflect the overall properties of populations (*29–31*). Indeed our results suggest that not only the gene content but also quantitative abundance and dynamics of individual traits should be considered when evaluating the ecological importance of differences among coexisting genotypes.

References and Notes

1. M. S. Rappé, S. J. Giovannoni, *Annu. Rev. Microbiol.* **57**, 369 (2003).
2. E. F. DeLong, *Curr. Opin. Microbiol.* **5**, 520 (2002).
3. T. Palys, L. K. Nakamura, F. M. Cohan, *Int. J. Syst. Bacteriol.* **47**, 1145 (1997).
4. S. F. Elena, R. E. Lenski, *Nat. Rev. Genet.* **4**, 457 (2003).
5. E. A. Joyce, K. Chan, N. R. Salama, S. Falkow, *Nat. Rev. Genet.* **3**, 462 (2002).
6. U. Bergthorsson, H. Ochman, *Mol. Biol. Evol.* **15**, 6 (1998).
7. J. C. Venter *et al.*, *Science* **304**, 66 (2004).
8. S. G. Acinas *et al.*, *Nature* **430**, 551 (2004).
9. V. Klepac-Ceraj *et al.*, *Environ. Microbiol.* **6**, 686 (2004).
10. J. R. Thompson *et al.*, *Appl. Environ. Microbiol.* **70**, 4103 (2004).
11. Materials and methods are available as supporting material on *Science* Online.
12. J. Thompson *et al.*, unpublished data.
13. J. B. Hughes, J. J. Hellmann, T. H. Ricketts, B. J. M. Bohannan, *Appl. Environ. Microbiol.* **67**, 4399 (2001).
14. N. Blackburn, T. Fenchel, *Mar. Ecol. Prog. Ser.* **189**, 1 (1999).
15. T. F. Thingstad, *Limnol. Oceanogr.* **45**, 1320 (2000).
16. R. A. Welch *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 17020 (2002).
17. R. A. Alm *et al.*, *Nature* **397**, 176 (1999).
18. J. A. Lindsay, M. T. G. Holden, *Trends Microbiol.* **12**, 378 (2004).
19. J. Hacker, E. Carniel, *EMBO Rep.* **2**, 376 (2001).
20. Y. Boucher *et al.*, *Annu. Rev. Genet.* **37**, 283 (2003).
21. J. G. Lawrence, H. Ochman, *Trends Microbiol.* **10**, 1 (2002).
22. H. Ochman, J. G. Lawrence, E. A. Groisman, *Nature* **405**, 299 (2000).
23. N. T. Perna *et al.*, *Infect. Immun.* **66**, 3810 (1998).
24. S. D. Reid, C. J. Herberlin, A. C. Bumbaugh, R. K. Selander, T. S. Whittam, *Nature* **406**, 64 (2000).
25. G. Rocap *et al.*, *Nature* **424**, 1042 (2003).
26. A. Mira, H. Ochman, N. A. Moran, *Trends Genet.* **17**, 589 (2001).
27. J. G. Lawrence, H. Ochman, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9413 (1998).
28. O. G. Berg, C. G. Kurland, *Mol. Biol. Evol.* **19**, 2265 (2002).
29. G. E. Fox, J. D. Wisotzkey, J. P. Jurtshuk, *Int. J. Syst. Bacteriol.* **42**, 166 (1992).
30. H. Sass, E. Wieringa, H. Cypionka, H. D. Babenzien, J. Overmann, *Arch. Microbiol.* **170**, 243 (1998).
31. E. Jaspers, J. Overmann, *Appl. Environ. Microbiol.* **70**, 4831 (2004).

32. We thank S. Chisholm, E. DeLong (both MIT), and F. Doolittle (Dalhousie University) for careful readings and comments on the manuscript; S. G. Acinas (MIT) for help with data analysis; J. Fontana and A. Stout (Massachusetts Department of Public Health) for instructions in PFGE optimization; and two anonymous reviewers for their insightful suggestions. This work was supported by grants from U.S. Department of Energy Genomes to Life program and NSF. Sequences have been submitted to GenBank under accession nos. AY836800 to AY837464.

Supporting Online Material

www.sciencemag.org/cgi/content/full/307/5713/1311/DC1
Materials and Methods
Fig. S1

5 October 2004; accepted 4 January 2005
10.1126/science.1106028

Optimization of Virulence Functions Through Glucosylation of *Shigella* LPS

Nicholas P. West,[1]*† Philippe Sansonetti,[2]* Joëlle Mounier,[2] Rachel M. Exley,[1] Claude Parsot,[2] Stéphanie Guadagnini,[3] Marie-Christine Prévost,[3] Ada Prochnicka-Chalufour,[4] Muriel Delepierre,[4] Myriam Tanguy,[2] Christoph M. Tang[1]‡

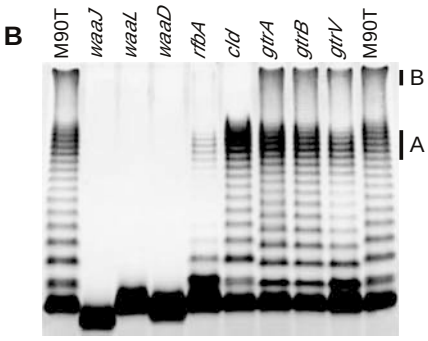
***Shigella*, the leading cause of bacillary dysentery, uses a type III secretion system (TTSS) to inject proteins into human cells, leading to bacterial invasion and a vigorous inflammatory response. The bacterium is protected against the response by the O antigen of lipopolysaccharide (LPS) on its surface. We show that bacteriophage-encoded glucosylation of *Shigella* O antigen, the basis of different serotypes, shortens the LPS molecule by around half. This enhances TTSS function without compromising the protective properties of the LPS. Thus, LPS glucosylation promotes bacterial invasion and evasion of innate immunity, which may have contributed to the emergence of serotype diversity in *Shigella*.**

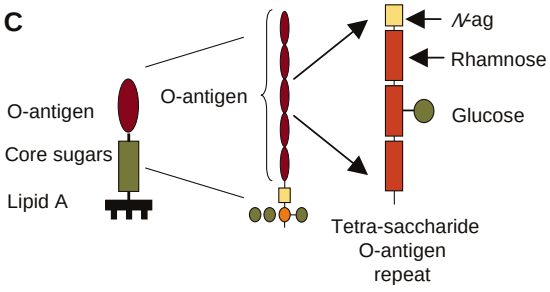
Pathogenic bacteria have evolved mechanisms to occupy specific niches within hosts while avoiding elimination by innate immune killing. *Shigella* is one of several pathogens that express a type III secretion system (TTSS), a needle-like structure

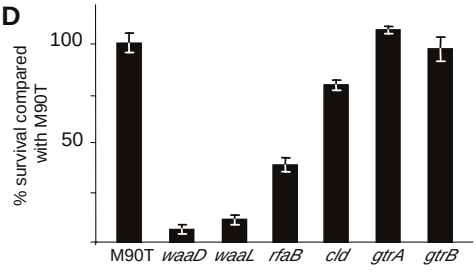
Fig. 1. (**A**) Genes disrupted in colonization-defective mutants. The competitive index (C.I.) assesses the ability of each mutant to colonize the rabbit ileal loop in direct comparison with the wild-type strain. A C.I. of 1 indicates no attenuation. (**B**) Tricine-SDS–polyacrylamide gel electrophoresis (TSDS-PAGE) of LPS from the mutants. Mode A and mode B LPS (indicated) contain around 15 and 100 subunits of O antigen, respectively. (**C**) Representation of *S. flexneri* 5a LPS (*N*-ag, *N*-acetylglucosamine). (**D**) Sensitivity of the colonization-defective mutants to the antimicrobial peptide, HBD-1; the percent survival relative to M90T after a 1-hour incubation is shown. Error bars indicate SEM.

A

Gene	Function	Competitive Index (C.I.)
waaJ	LPS core biosynthesis	0.08
waaD	LPS core biosynthesis	0.10
waaL	O-antigen addition	0.10
rfbA	dTDP rhamnose production	0.10
cld	Chain length determinant	0.04
gtrA	O-antigen glucosylation	0.02
gtrB	O-antigen glucosylation	0.03







through which effector molecules are translocated from the bacterium directly into human cells (*1*). This molecular syringe is required by *Shigella* for entry into epithelial cells (*2*–*4*). Bacterial invasion is critical for gastrointestinal (GI) disease because it elicits an intense inflammatory response with subsequent disruption of the epithelial barrier and formation of abscesses in the mucosa. Thus, *Shigella* both inflicts damage to host tissues and survives the vigorous innate immune response. We used signature-tagged mutagenesis (*5*) to understand the genetic basis of GI infection by the *Shigella flexneri* serotype 5 strain, M90T. Out of 2900 mutants analyzed in the rabbit ligated ileal loop model of shigellosis (*6*, *7*), 15 colonization-defective mutants were identified, with 7 resulting from loss of genes involved in lipopolysaccharide (LPS) biosynthesis (Fig. 1, A and B).

LPS of enteric bacteria consists of three regions: lipid A, an inner core, and repeating O-antigen subunits. The *Shigella* O-antigen subunit is a tri-rhamnose (rha)-*N*-acetyl glucosamine (*N*-ag) tetrasaccharide (*8*), modified by the addition of glucose and/or acetate, depending on the serotype (Fig. 1C). The modal number of subunits is about 15 (mode A, Fig. 1B), with another prevalent population of around 100 repeats (mode B). The *waaJ*, *waaD*, and *waaL* mutants, which cannot synthesize the inner core or ligate the O antigen to the inner core, were markedly attenuated (Fig. 1A), which probably reflects their failure to resist being killed by antimicrobial peptides expressed in the GI tract (*9*) (Fig. 1D; fig. S1). This probably also accounts for the attenuation of the *rfbA* mutant, which fails to convert glucose-1-phosphate to deoxythymidine diphosphate–rhamnose (dTDP-rha), a step required for the efficient synthesis of O antigen. Likewise the *cld* mutant has a truncated LPS molecule lacking mode B of the O antigen (Fig. 1B) (*10*). The colonization capacity of these mutants was restored by complementation with the corresponding wild-type allele (table S1 and fig. S2).

Two mutants were isolated with insertions in the O-antigen glucosylation operon. This operon consists of *gtrA*, *gtrB*, and *gtrV* present on a resident bacteriophage (*11*, *12*), and confers serotype-specific glucosylation of LPS. The *gtrA* and *gtrB* mutants had subtle alterations in their LPS profile and expressed a number of O-antigen subunits equivalent to the wild-type strain. However, each subunit had a slightly reduced molecular mass (Fig. 1, B and C) through lack of glucosylation on the second rhamnose residue (rha^{II}) of the O antigen (*13*). To confirm the role of the *gtr* operon in colonization, we constructed a *gtrV* deletion mutant. *gtrV* is predicted to encode the glucosyltransferase, which adds glucose to the O antigen in a serotype-specific manner (*14*). The strain lacking *gtrV* (M90Δ*gtrV*) had a substantial disadvantage for survival in the GI tract (Fig. 2A). Complementation of each *gtr* mutant with pNW83, which contains all *gtr* genes (Fig. 2), demonstrated that expression of serotype 5a LPS confers a marked survival advantage compared with isogenic strains expressing unglucosylated LPS.

To establish whether O-antigen modifications of other *Shigella* serotypes also confer a competitive advantage, we constructed isogenic strains expressing LPS from different serotypes (Fig. 2B). The *gtr* operon was deleted from M90T (generating M90TΔ*gtr*), then the *gtr* operons from *S. flexneri* 1a, 2a, and 5a (*13*) were introduced on plasmids resulting in M90TΔ*gtr*p1a, M90TΔ*gtr*p2a, and M90TΔ*gtr*p5a, respectively (Fig. 2B). Furthermore, the entire *gtr* operon from *S. flexneri* 2a was introduced chromosomally into M90TΔ*gtr*, resulting in M90T2a. The serotype replacements were confirmed by agglutination assays and LPS profiling (table S2 and Fig. 2C). The O-antigen subunits of M90TΔ*gtr*p5a were more extensively glucosylated than those in M90T, probably because of a gene dosage effect. Glucosylation of the O antigen restored the virulence of *S. flexneri* in the GI tract (competitive index of 0.88, 0.92, and 1.1 for M90TΔ*gtr*p1a, M90TΔ*gtr*p2a, and M90TΔ*gtr*p5a, respectively), which demonstrates that LPS glucosylation promotes the fitness of *Shigella* in vivo, irrespective of serotype.

One potential explanation for the attenuation of the *gtr* mutants is that they are de-

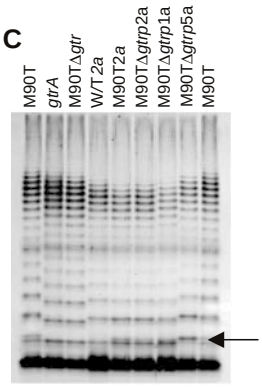
[1]Centre for Molecular Microbiology and Infection, Department of Infectious Diseases, Faculty of Medicine, Flowers Building, Imperial College London, London SW7 2AZ, UK [2]Unité de Pathogénie Microbienne Moléculaire, Unité INSERM 389; [3]Plate-Forme de Microscopie Electronique; and [4]Unité de Résonance Magnétique Nucléaire des Biomolécules, URA 2185 CNRS; Unité INSERM 389, Institut Pasteur, 28 rue du Dr. Roux, F-75724 Paris Cédex 15, France.

*These authors contributed equally to this work.
†Present address: Centenary Institute of Cancer Medicine and Cell Biology, Newtown, NSW 2042, Australia.
‡To whom correspondence should be addressed. E-mail: c.tang@imperial.ac.uk

A

Strain	C.I.
M90TΔ*gtrV*	0.045
M90TΔ*gtrV* pNW83	1
M90TΔ*gtrA* pNW83	1
M90TΔ*gtrB* pNW83	1
M90TΔ*gtrA* vs M90T	0.07
BS176Δ*gtrA* vs BS176	0.55

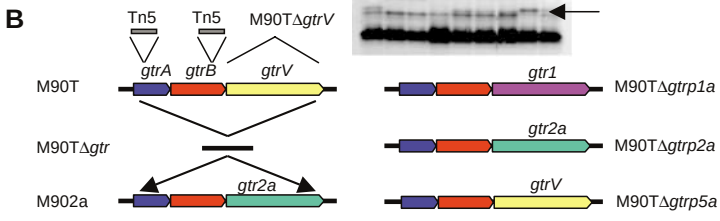
C



E

Strain	No. bacteria
M90T	1.89 (0.01)
M90TΔ*mxiD*	0.02 (0.00)
M90TΔ*gtr*	0.23 (0.01)
M90TΔ*gtrp5a*	7.3 (2.6)
M90TΔ*cld*	4 (1.6)
M90TΔ*galU*	11.3 (0.33)

B



D

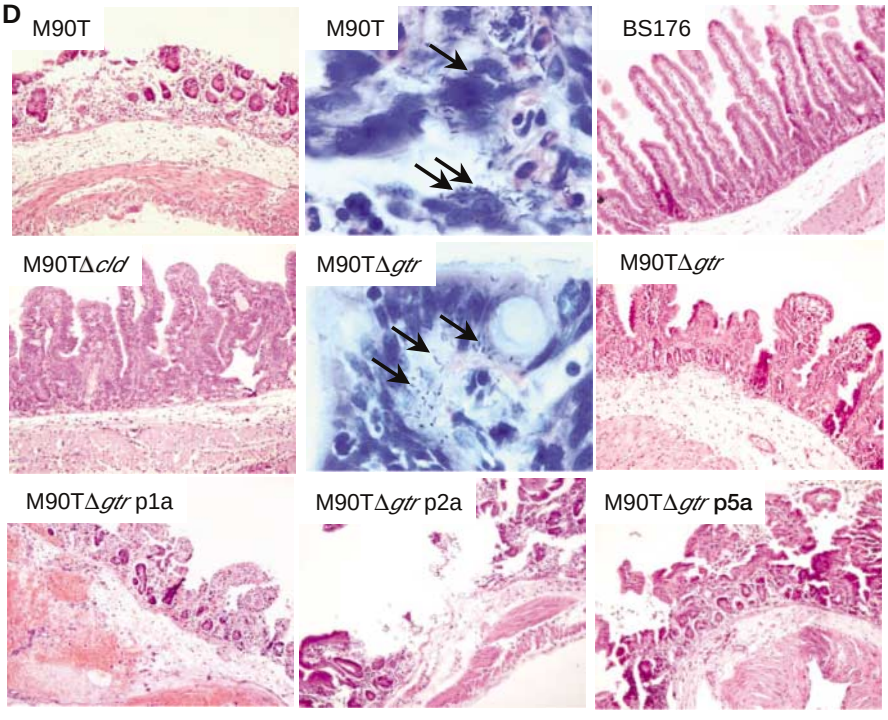


Fig. 2. (**A**) Competitive index of mutants with defects in genes of the *gtr* operon. (**B**) Construction of strains expressing LPS from different serotypes. (**C**) TSDS-PAGE analysis of LPS from various strains. The arrow indicates enhanced glucosylation of M90TΔ*gtr*p5a compared with M90T. (**D**) Histological analysis of infected ileal loops stained with hematoxylin, eosin, and safranin or with Evans blue to demonstrate the location of bacteria (indicated by arrows). The infecting strain is shown. (**E**) The invasive capacity of strains expressing LPS from different serotypes. The average number of invasive bacteria per epithelial cell is indicated (SD in parentheses).

fective in avoiding host clearance by innate immune mechanisms. However, the *gtr* mutants were fully able to withstand the adverse effects of bile salts, hyperosmolarity [to 300 mosmol (*15*)], complement-mediated lysis (fig. S3), and antimicrobial peptides (Fig. 1D; fig. S1). Furthermore, the survival disadvantage resulting from loss of *gtrA* was not evident in a noninvasive strain of *S. flexneri*, BS176, which lacks the TTSS (Fig. 2A). Thus, the decreased fitness of the *gtr* mutants was only manifest after mucosal invasion.

We next investigated whether LPS glucosylation contributes to the proinflammatory host response, the hallmark of shigellosis. Infection of ileal loops with M90T led to dramatic alterations of mucosal tissues with rupture and destruction of the intestinal epithelium (Fig. 2D). The lesions, typical of acute shigellosis, included extensive zones of epithelial detachment and loss of villi. A massive polymorphonuclear leukocyte (PMN) response was present within the lamina propria and in the edematous submucosal tissues. Numerous invasive bacteria were seen in abscesses and in the lamina propria. Infection with the noninvasive isolate, BS176, did not result in significant alteration of the mucosa, and M90TΔ*cld* (which expresses a truncated O antigen) also had attenuated virulence (Fig. 2D). M90TΔ*gtr* caused consistently less pathological damage than M90T. Intestinal villi were only shortened and swollen, and zones of epithelial destruction were restricted, with abscesses limited to the base of villi (Fig. 2D). The number of bacteria and infiltrating PMNs was much lower than observed in M90T-infected tissues. Total restoration of the virulent phenotype was observed after infection with M90TΔ*gtr* expressing a glucosyltransferase operon, irrespective of serotype (Fig. 2D).

We hypothesized that LPS glucosylation contributes to the virulence of *S. flexneri* through an effect on its invasive potential. Consistent with this, each *gtr* mutant had a substantially reduced ability to invade epithelial cells when compared with M90T (Fig. 2E); this defect was entirely corrected by reintroducing the *gtr* operon from any serotype. Furthermore, strains with increased LPS glucosylation (e.g., M90TΔ*gtr*p5a) showed enhanced invasion compared with M90T (Fig. 2E). We considered whether changes in surface hydrophobicity of the strains could explain this finding. The increased hydrophobicity of M90TΔ*galU*, which expresses a truncated LPS, might be responsible for the increased adherence of bacteria to eukaryotic cells (table S3). However, the slight variations in hydrophobicity among M90T, M90TΔ*cld*, M90TΔ*gtr*, and the complemented strain could not account for the differences in invasion (table S3). Furthermore, the initial adhesion of the *gtr* strains to epithelial cells was not affected (table S4), which indicated that loss of O-antigen glucosylation did not affect exposure of any tissue-specific adhesin.

Three-dimensional molecular models of serotype 5a of the O antigen based on nuclear magnetic resonance (NMR) data give structures adopting a right-handed, threefold helix with the branched glucosyl residues pointing outward (*16*). Glucosylation of serotype 5a LPS is predicted to induce a transition from a linear to helical conformation with the glucosyl residue exposed on the exterior of the helix, forming a more compact structure (Fig. 3A) than unglucosylated LPS (*17*). This would dramatically shorten the O antigen, halving the distance it extends beyond the outer membrane. In the mode A LPS, the O antigen would extend around 21 nm from the outer membrane in the absence of glucosyla-

tion, but only 11 nm for a glucosylated O antigen. To test these predictions, bacteria were visualized by transmission electron microscopy (Fig. 3). M90T had a dense surface material extending about 35 nm beyond the outer membrane (between the two arrowheads, Fig. 3B). In contrast, the exterior of M90TΔ*gtr* was composed of more diffuse, filamentous material that extended around 70 nm from the outer membrane (Fig. 3B). The surface of M90TΔ*gtr*p5a, which displays enhanced glucosylation, was even more compact than M90T. As the average length of protruding needles of the *Shigella* TTSS is 60 nm (*18*), LPS glucosylation might affect TTSS function, which would account for the changes in cell invasion (Fig. 2E).

To establish whether LPS glucosylation affects the exposure of TTSS on the cell surface, we examined bacteria by scanning electron microscopy (Fig. 4A). TTSS needles were seen clearly on M90T but were absent in M90TΔ*mxiD*. A significantly lower number of TTSS needles were detected on M90TΔ*gtr* compared with M90T, and the defect was entirely restored when the *gtr* operon was introduced (Fig. 4B). To confirm that these structures are TTSS needles, we performed immunolabeling with a monoclonal antibody against IpaB, which is both secreted and exposed at the TTSS tip (*19–21*). This revealed surface structures on M90T that were much less frequently seen in M90TΔ*gtr* and were absent in the TTSS-null mutant. This reduced staining was not due to impaired secretion, because secretion of IpaB and IpaC was identical in M90T and M90TΔ*gtr* (Fig. 4D). Furthermore with negative-staining electron microscopy, similar numbers of TTSS needles were observed emerging from the outer membrane of M90T and M90TΔ*gtr* (Fig. 4E), which shows that glucosylation does not affect the assembly of TTSS in the outer membrane.

Our results show that LPS dictates the key function of the *Shigella* TTSS of mediating bacterial invasion into host cells. Strains with truncated LPS are highly proficient at invading cells in vitro, possibly through enhanced access of the TTSS to host cells and altered hydrophobicity (Fig. 4E). However, this advantage is entirely offset in vivo where the bacterium is more susceptible to innate immune factors. In strains expressing full-length LPS in the absence of glucosylation, although *Shigella* is fully able to resist innate immune killing, the extended LPS isoform impairs TTSS function, which leads to reduced virulence within the GI tract. In both instances, the bacterium is at a competitive disadvantage through a trade-off between its invasive capacity and its ability to survive innate immune effectors.

Evidently, glucosylation of LPS facilitates invasion of target cells by altering the conformation of LPS to optimize the exposure of TTSS needles while retaining resistance against host defences. The relationship between LPS and the function of the TTSS can be considered similar to a sword and shield, in which there is a balance between the length of LPS O side chains to protect the bacterium against innate immune effectors ("the shield"), and the influence on the function of the TTSS needle ("the sword"): both are essential for *Shigella* virulence.

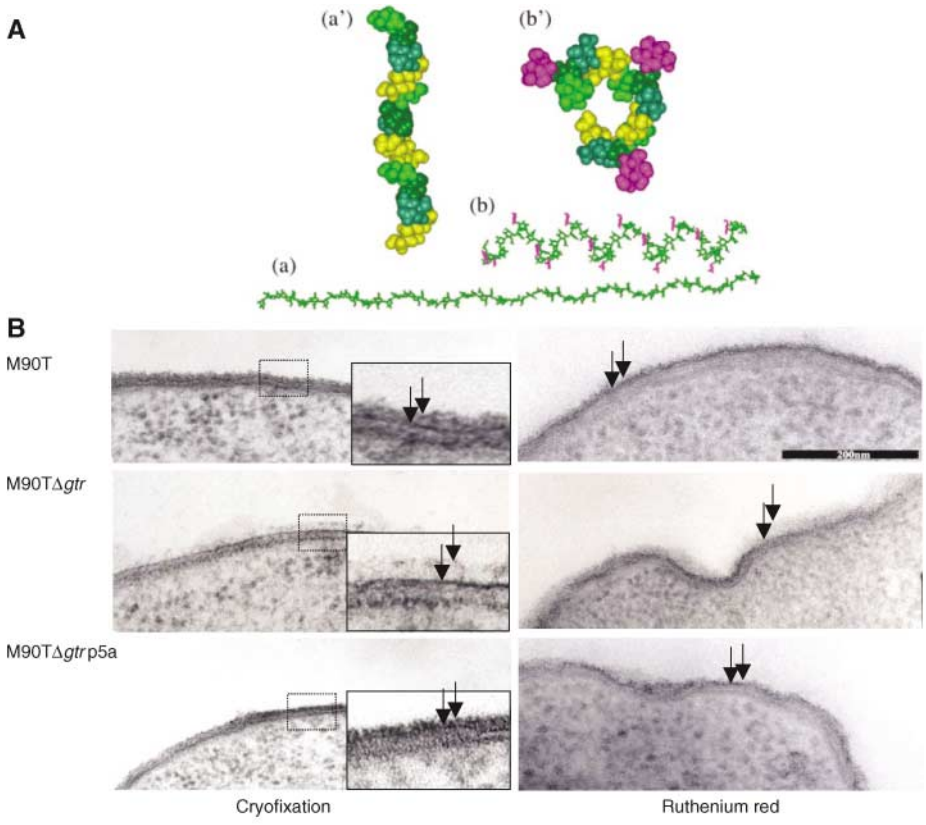


Fig. 3. (**A**) Models of *S. flexneri* unglucosylated (a) and glucosylated (b) O antigens formed by 15 repeating units depicted with branched glucosyl residues colored in magenta. The distances between the C1 atoms of the first and last subunits are 21.2 nm and 11.3 nm, for unglucosylated and glucosylated O antigens, respectively. View of space-filling models, a′ and b′ (shades of green, rhamnopyranosyl residues; yellow, glucopyranosyl). (**B**) Surface topology of strains revealed by transmission electron microscopy after cryofixation or treatment with ruthenium red. The electron-dense material at the bacterial surface is indicated between the arrows. In M90TΔ*gtr*, the surface material is less compact (reaching about 70 nm beyond the outer membrane) than M90T (extending around 35 nm). Surface staining is more intense and condensed on M90TΔ*gtr*p5*a* than on M90T. The regions within boxes have been expanded.

References and Notes

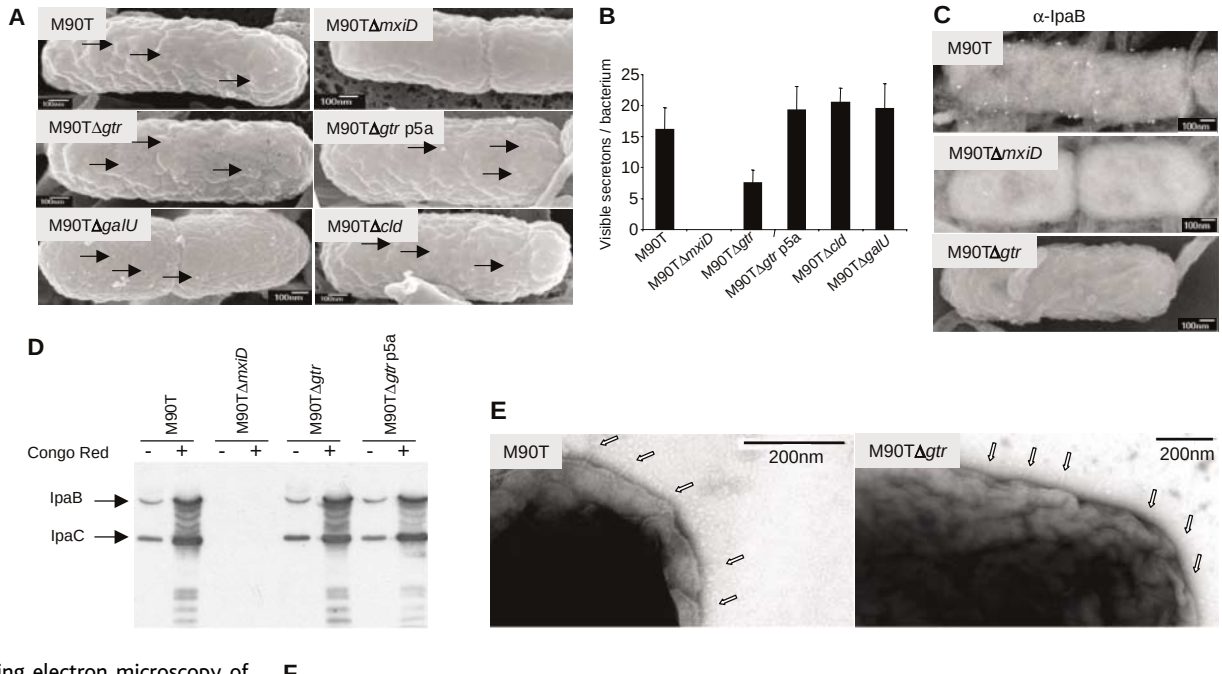
1. B. B. Finlay, P. Cossart, *Science* **276**, 718 (1997).
2. P. J. Sansonetti, D. J. Kopecko, S. B. Formal, *Infect. Immun.* **35**, 852 (1982).
3. A. T. Maurelli, B. Baudry, H. d'Hauteville, T. L. Hale, P. J. Sansonetti, *Infect. Immun.* **49**, 164 (1985).
4. C. Buchrieser *et al.*, *Mol. Microbiol.* **38**, 760 (2000).
5. M. Hensel *et al.*, *Science* **269**, 400 (1995).
6. P. J. Sansonetti, J. Arondel, J. M. Cavaillon, M. Huerre, *J. Clin. Invest.* **96**, 884 (1995).
7. Materials and Methods are available as supporting online material.
8. L. Kenne, B. Lindberg, K. Petersson, E. Katzenellenbogen, E. Romanowska, *Eur. J. Biochem.* **91**, 279 (1978).
9. R. N. Cunliffe, Y. R. Mahida, *J. Leukoc. Biol.* **75**, 49 (2003).
10. M. Hong, S. M. Payne, *Mol. Microbiol.* **24**, 779 (1997).
11. P. T. Huan, D. A. Bastin, B. L. Whittle, A. A. Lindberg, N. K. Verma, *Gene* **195**, 217 (1997).
12. D. A. Simmons, *Biochem. Soc. Trans.* **18**, 1271 (1990).
13. G. E. Allison, N. K. Verma, *Trends Microbiol.* **8**, 17 (2000).
14. S. Guan, D. A. Bastin, N. K. Verma, *Microbiology* **145**, 1263 (1999).
15. N. P. West, P. Sansonetti, C. M. Tang, unpublished observations.
16. M.-J. Clément *et al.*, *J. Biol. Chem.* **278**, 47928 (2003).
17. K. Bock *et al.*, *J. Chem. Soc., Perkin Trans.* **2**, 59 (1982).
18. A. Blocker *et al.*, *J. Cell Biol.* **147**, 683 (1999).
19. K. Ito, T. Nakajima, T. Sasaki, H. Watanabe, *Microbiol. Immunol.* **35**, 335 (1991).
20. M. M. Venkatesan, J. M. Buysse, E. V. Oaks, *J. Bacteriol.* **174**, 1990 (1992).
21. A. Skoudy *et al.*, *Cell. Microbiol.* **2**, 19 (2000).

22. The work was supported by the European Union (QLRT-1999-00938), and an International Scholarship from the Howard Hughes Medical Institute to P.S. We thank D. W. Holden, S. Marshall, E. R. Moxon, S. Keshav, and H. Smith for their invaluable contributions.

Supporting Online Material

www.sciencemag.org/cgi/content/full/307/5713/1313/DC1
Materials and Methods
Figs. S1 to S3
Tables S1 to S6
References and Notes

8 December 2004; accepted 14 January 2005
10.1126/science.1108472



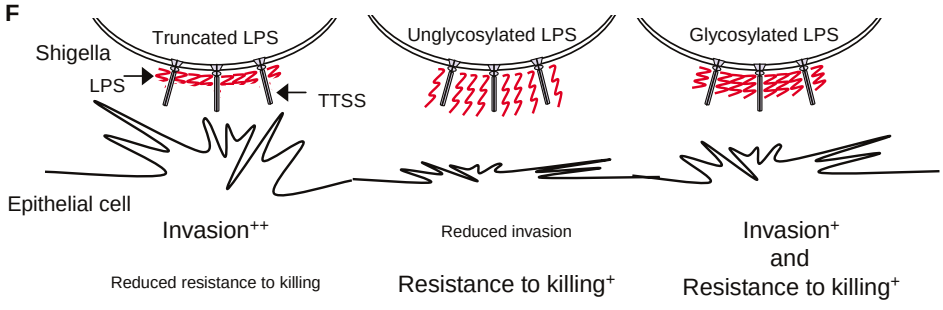


Fig. 4. (**A**) Scanning electron microscopy of bacteria demonstrating the abundance of TTSS secretions (arrowed) on the cell surface; scale bars and strains are shown. (**B**) Number of visible secretions per bacterium (error bars show SD). (**C**) Immunogold labeling of IpaB at the bacterial surface, and (**D**) secretion of IpaB and IpaC into culture supernatants detected by Coommassie are not affected by LPS glucosylation; secretion was induced by the addition of Congo Red (+). (**E**) Negative-staining electron microscopy showed that the number of TTSS emerging from the bacterial outer membrane was not altered in the strain lacking the *gtr* operon. (**F**) The interaction between LPS and TTSS. Hyperinvasive strains that have truncated LPS molecules are susceptible to being killed by the innate immune response in vivo. Bacteria expressing unglucosylated LPS are compromised for invasion and therefore attenuated. Glucosylation of the O antigen halves the length of the LPS molecule, which allows efficient function of the TTSS while it retains resistance to antimicrobial factors in vivo.

NMR Structure of Mistic, a Membrane-Integrating Protein for Membrane Protein Expression

Tarmo P. Roosild, Jason Greenwald, Mark Vega, Samantha Castronovo, Roland Riek,* Senyon Choe*

Although structure determination of soluble proteins has become routine, our understanding of membrane proteins has been limited by experimental bottlenecks in obtaining both sufficient yields of protein and ordered crystals. Mistic is an unusual *Bacillus subtilis* integral membrane protein that folds autonomously into the membrane, bypassing the cellular translocon machinery. Using paramagnetic probes, we determined by nuclear magnetic resonance (NMR) spectroscopy that the protein forms a helical bundle with a surprisingly polar lipid-facing surface. Additional experiments suggest that Mistic can be used for high-level production of other membrane proteins in their native conformations, including many eukaryotic proteins that have previously been intractable to bacterial expression.

Integral membrane (IM) proteins, constituting nearly 30% of eukaryotic genomes, play central roles in cellular transport processes, intercellular signaling, and growth regulation. However, of the more than 28,000 high-resolution protein structures known, only some 25 unique families of IM proteins are represented. This disparity is accounted for by two bottlenecks in membrane protein structural analysis: high-yield protein production and crystallization. Recombinant expression of IM proteins in *Escherichia coli*, the primary protein source for biophysical studies, has met with limited success (*1*). Two complications likely account for this difficulty. First, IM proteins must be trafficked to the membrane, requiring targeting signals that may not be recognized by the bacterial host. Second, high-level expression of membrane proteins that can use *E. coli* translocon machinery will competitively exclude production of other vital host membrane proteins, leading to toxicity. Most successful attempts at expression of IM proteins in bacteria have used low-copy-number plasmids with weak promoters to produce low levels of protein, compensated by large culture volumes (*2*).

Structural Biology Laboratory, Salk Institute, San Diego, CA 92037, USA.

*To whom correspondence should be addressed. E-mail: choe@salk.edu, riek@salk.edu

Alternatively, one can target IM proteins to inclusion bodies (*3*), but this requires subsequent renaturation of the desired protein from these insoluble deposits, a process with limited success rates.

The established procedure of using fusion partner proteins to aid production of recombinant proteins has also had limited utility in the production of eukaryotic IM proteins (*4*), because the fusion proteins currently available do not target the construct to the membrane or facilitate membrane insertion. An ideal fusion partner for IM protein production would autonomously traffic its cargo to the membrane, bypassing the translocon and associated toxicity issues while retaining the characteristics of other successful fusion partner proteins, including relatively small size, in vivo folding, and high stability. Several proteins (particularly bacterial toxins) and some synthetic peptides (*5*) have many of these characteristics, which suggests that an ideal fusion partner specialized for recombinant IM protein production in *E. coli* is likely to exist.

Crystallization is another obstacle in the determination of IM protein structures, because such proteins must be solubilized in detergent micelles that are inherently resistant to forming ordered crystal lattices. Nuclear magnetic resonance (NMR) spectroscopy offers an alternative method for determining atomic resolution structures of proteins (*6*, *7*). To date, however, protocols for NMR structure determination of IM proteins have been established only for very small, structurally simplistic IM proteins (*8*, *9*) and for outer membrane bacterial porins (*10*–*12*), whose β-barrel fold allows collection of ample interstrand long-range backbone-backbone nuclear Overhauser effects (NOEs) that are sufficient to determine the fold of the protein. The development of new techniques to specifically address the inherent characteristics of α-helical IM proteins is necessary to bring the powerful tools of NMR to bear on this class of molecules.

We have isolated a 110–amino acid (13 kD) *B. subtilis* protein called Mistic (an acronym for "membrane-integrating sequence for translation of IM protein constructs"). Mistic associates tightly with the bacterial membrane when expressed recombinantly in *E. coli* (Fig. 1A) (*13*). Surprisingly, however, Mistic is highly hydrophilic, lacking a recognizable signal sequence. Detergent-solubilized Mistic binds tightly to micelles and aggregates rapidly when stripped of surfactant. Mistic solubilized in lauryl dimethylamine oxide (LDAO) was found to be monomeric by static light scattering analysis in combination with detection of ultraviolet absorption (fig. S1), forming a protein-detergent complex (PDC) of ~25 kD containing ~50 molecules of LDAO (relative molecular mass = 229.4) per molecule of Mistic.

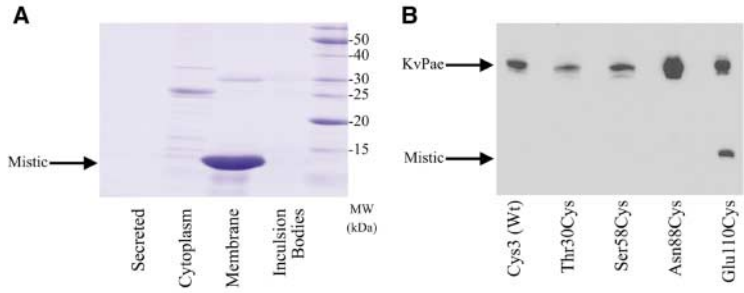


Fig. 1. Mistic characterization. (**A**) SDS–polyacrylamide gel electrophoresis (PAGE) results for Ni–nitrilotriacetic acid (NTA) elutions from fractionation of a culture recombinantly expressing octahistidine-tagged Mistic. Mistic is found abundantly only in the bacterial membrane. (**B**) Topology analysis of Mistic as assessed by biotinylation of monocysteine variations of Mistic by the membrane-impermeable, thiol-reactive probe MPB. Only Glu^{110} at the C terminus is well exposed periplasmically. Cys^{3} at the N terminus of the protein and the centrally located Ser^{58}, both also putatively on the extracellular side of the membrane, are nonreactive with MPB in right-side-out (RSO) membrane vesicle preparations, consistent with these side chains being embedded in the membrane. In support of this hypothesis, Cys^{3} mutation to Ser is functionally disruptive, whereas mutation to hydrophobic Val, Leu, or Ile is well tolerated. Mistic constructs were expressed as a fusion to a bacterial potassium channel (KvPae) and subsequently separated by cleavage with thrombin. The channel, identical in all constructs, serves as an internal control for calibrating expression, extraction, biotinylation, and detection efficiency among the samples.

The in vivo topology of this protein in *E. coli* was analyzed by evaluating the accessibility of an array of monocysteine mutants to the membrane-impermeable thiol biotinylating reagent 3-(*N*-maleimido-propinyl) biocytin (MPB) (*14*). In addition to the single naturally occurring cysteine (residue 3), cysteine mutations were introduced individually at the C terminus (residue 110) and in predicted loop regions at positions 30, 58, and 88 (Fig. 2A), with the naturally occurring cysteine mutated to valine. This experiment revealed a well-exposed periplasmic C terminus (Fig. 1B) (fig. S2). The lack of reactivity of the other locations indicates that they are either intracellular or membrane-embedded in Mistic's native conformation.

NMR de novo structure determination began with sequential backbone assignment, including the use of transverse relaxation optimized spectroscopy (TROSY)–HNCA (*15*, *16*), TROSY-$HNCA^{coded}CO$ (*17*), and TROSY-based ^{15}N-resolved [^{1}H,^{1}H]–nuclear Overhauser effect spectroscopy (NOESY) (mixing time 200 ms) of a ^{2}H, ^{15}N, and ^{13}C-labeled sample (fig. S3). The $^{13}C^{\alpha}$ chemical shift deviation from "random coil" values, the observed NOE pattern, and slow $^{1}H_{N}$ exchange with solvent strongly indicate the presence of four helices comprising residues 8 to 22, 32 to 55, 67 to 81, and 89 to 102 (Fig. 2A). Although intraresidue, sequential, and medium-range NOEs and angle restraints enabled the assignment of secondary structure, without long-range restraints the fold of the protein could not be determined. We thus used the monocysteine mutant library described in the topology assay (see above) to incorporate site-directed spin labels within Mistic that produce distance-dependent line-broadening perturbations in the NMR spectra (*18*) that could be translated into distances for structure determination (*19*). [^{15}N,^{1}H]-TROSY experiments were measured on Mistic samples modified with the thiol-reactive nitroxide label (1-oxyl-2,2,5,5-tetramethyl-Δ^{3}-pyrroline-3-methyl) methanethiosulfonate (MTSL) (Fig. 2B). The signal changes observed for the five spin-labeled samples were transformed into 197 long-range upper-distance and 290 lower-distance restraints (fig. S4).

Initial structure calculation was performed with CYANA (*20*) using the collected NOE data, chemical shift–derived angle restraints, and restraints derived from spin labeling. In addition, α-helical hydrogen bond restraints were implemented for residues that show all of the three following properties: slow HN exchange, a helical ^{13}C chemical shift, and helical backbone NOEs (Fig. 2A). In an iterative process, the derived scaffold was used to collect long-range and medium-range NOEs and to refine calibration of the spin-label restraints. In the end, 29 long-range NOEs between methyl or aromatic protons and amide protons were identified. Because these distances are intrinsically large in a helical bundle and concomitantly result in weak NOEs, the use of a cryoprobe and long mixing times of 200 ms were essential.

The final structure calculation was performed with 573 NOE distance restraints, 346 angle restraints from chemical shifts and NOEs, and 478 distance restraints from the spin-label experiments (table S1). A total of 100 conformers were initially generated by CYANA; in Fig. 2C, the bundle of 10 conformers with the lowest target function is used to represent the three-dimensional NMR structure. The resulting structure is a four-helix bundle (Fig. 3A). Although all helices except α2 are slightly shorter (~14 amino acids) than expected for a bilayer-traversing helix, this is likely due to partial unraveling of the ends of the helices in the detergent micelle environment, especially at the N and C termini (α1 and α4). Helix α2 has a kink,

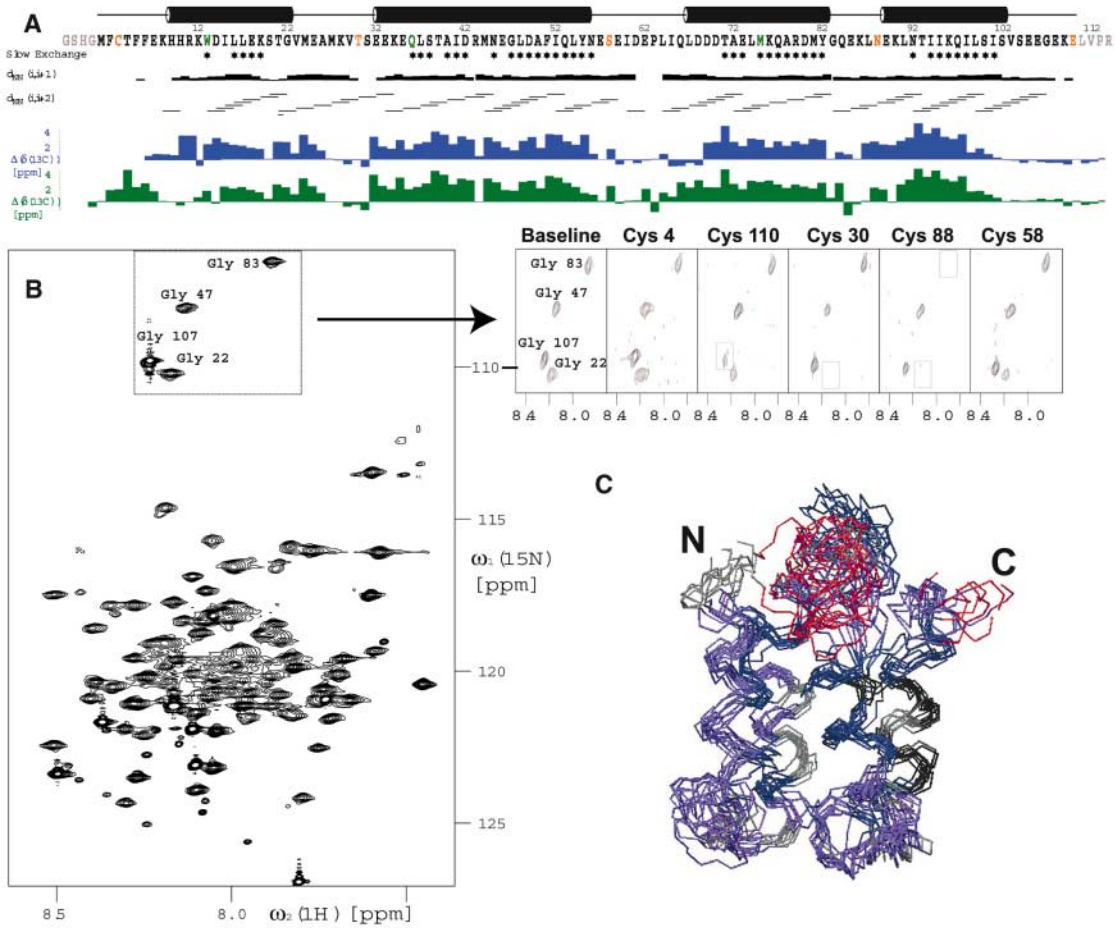


Fig. 2. Secondary structure and long-range interactions of Mistic. (**A**) Primary sequence of Mistic displaying location of monocysteine probing residues (orange), structural disruption mutants (green), and cloning artifact residues (gray) with secondary structural boundaries above the sequence. (First line) $^{1}H_{N}$ protection from solvent exchange indicative for hydrogen bond formation (stars). The solvent protection is determined by the absence of a cross-peak between the chemical shifts of $^{1}H_{N}$ and water in the ^{15}N-resolved TROSY-[^{1}H,^{1}H]-NOESY spectrum. (Second and third lines) NOEs observed in the ^{15}N-resolved TROSY-[^{1}H,^{1}H]-NOESY. Thin, medium, and thick bars represent weak (4.5 to 5.5 Å), medium (3 to 4.5 Å), and strong (< 3 Å) sequential NOEs [$d_{NN}(i, i + 1)$]. The medium-range NOEs [$d_{NN}(i, i + 2)$] are shown by lines starting and ending at the positions of the residues related by the NOE. (Fourth and fifth lines) Deviation of the $^{13}C^{\alpha}$ chemical shifts from corresponding "random coil" chemical shifts in 0 mM K^{+} (blue) and 100 mM K^{+} (green), as independently assigned. Values larger than 1.5 ppm are indicative of an α-helical secondary structure; values smaller than −1.5 ppm are indicative of β-sheet secondary structure. (**B**) The 2D [^{15}N,^{1}H]-TROSY spectrum of Mistic is shown along with parts of the 2D [^{15}N,^{1}H]-TROSY spectra in the presence of paramagnetic spin labels at positions Cys^{3}, Thr^{30}Cys, Ser^{58}Cys, Asn^{88}Cys, and Glu^{110}Cys. Comparison of peaks' heights between perturbed spectra and multiple reference spectra was used to obtain long-range distance restraints. (**C**) Superposition of 10 conformers representing the final NMR structure. The bundle is obtained by superimposing the backbone C^{α} carbons of residues 13 to 62 and 67 to 102. The bundle is colored by ^{15}N{^{1}H}NOE data by the following color code: black, 1 to 0.8; navy, 0.8 to 0.6; blue, 0.6 to 0.4; red, 0.4 to 0.2. ^{15}N{^{1}H}NOE as well as T1 (^{15}N) and T2 (^{15}N) relaxation data indicate that the dynamics of the structure is generally reflected in the variance of the conformers. In particular, the loop connecting α2 and α3, as well as the C terminus of Mistic, are more mobile. The T1/T2 ratio of ^{15}N was used to estimate the effective global rotational correlation time at 11 ns. This value corresponds to a spherical molecule of ~22 kD.

centrally positioned and putatively within the membrane. Most surprising, Mistic retains an unexpectedly hydrophilic surface for an IM protein even though it is assembled internally with a typical hydrophobic core (Fig. 3, B and C).

Given the membrane-traversing topology demonstrated by the MPB labeling experiment (Fig. 1B), this unusual surface property is very intriguing. To confirm the orientation of Mistic with respect to the membrane, we measured and assigned NOEs between Mistic and its solubilizing LDAO detergent micelle. When sites with NOE signals are mapped to the surface of the Mistic structure, a concentric ring of detergent interactions around the helical bundle is observed, as expected for a membrane-integrated protein (Fig. 4, A to C). Additionally, we perturbed Mistic spectra with paramagnetic probes that selectively partition to hydrophilic or hydrophobic environments (fig. S5) (*21*). The results from this study are also consistent with Mistic being embedded within the LDAO micelle.

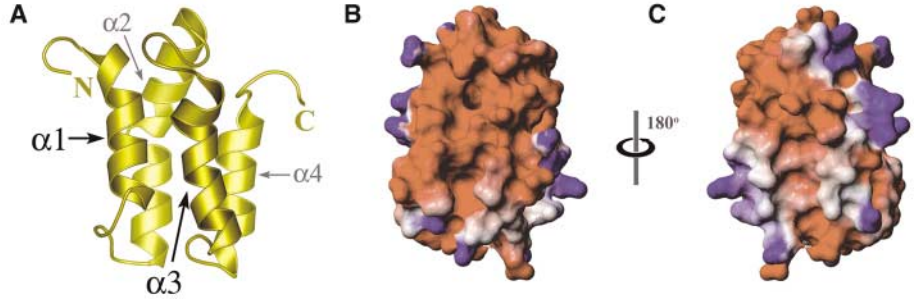


Fig. 3. Mistic structure. (**A**) Ribbon diagram of the lowest energy conformer highlighting the four α-helix bundle. (**B**) Surface representation of Mistic, oriented as in (A), mapping electrostatic potential. Color code is blue for positive charges, red for negative charges, and white for neutral surface. (**C**) Electrostatic potential of Mistic, viewed from the opposite face from that shown in (B).

We hypothesized that Mistic might be exploited to target another protein to the bacterial membrane, when fused to Mistic's C terminus, such that it too could readily fold into its native, lipid bilayer–inserted conformation. We tested the Mistic-assisted expression of three topologically and structurally distinct classes of eukaryotic IM proteins: voltage-gated K^{+} chan-

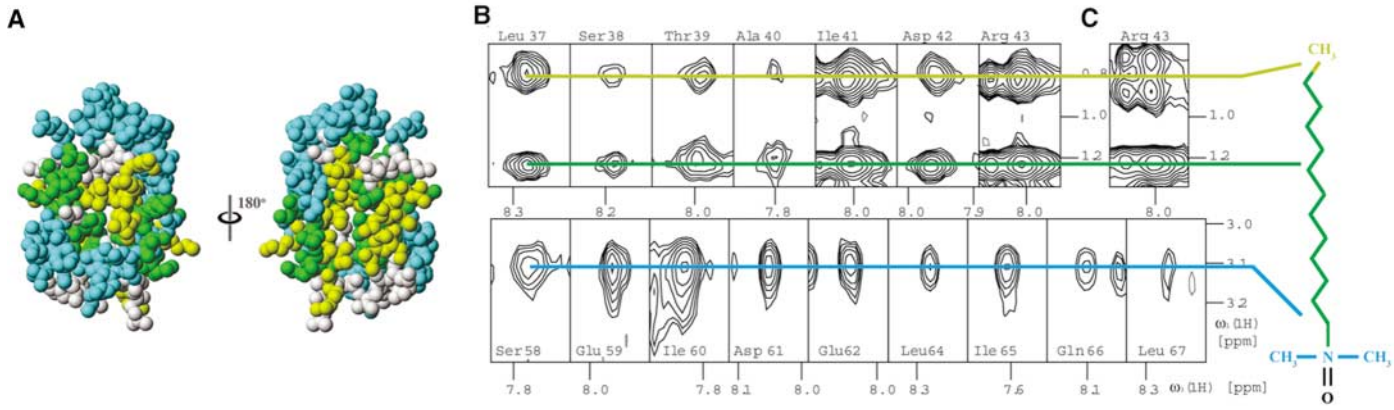


Fig. 4. Mistic-detergent interactions. (**A**) Surface representation of Mistic indicating observed NOE interactions between detergent molecules and the protein. Observed interactions are coded blue between the head methyl (CH_3) groups of LDAO and backbone amides (^1H_N) of the protein, yellow between the hydrophobic CH_3 end of LDAO and ^1H_N, and green between the LDAO chain (CH_2) and ^1H_N. NOEs were never observed from the same residue to both the head-group methyl and the aliphatic chain tail methyl of LDAO for the same residue. (**B**) A selection of intermolecular NOEs between LDAO and residues 37 to 43 and 58 to 67 of Mistic; [^{15}N,^1H] strips from the ^{15}N-resolved TROSY [^1H,^1H]-NOESY are shown. The detergent-protein NOEs are marked by a bar colored as in (A), pointing to the appropriate portion of the chemical structure of LDAO. (**C**) For the differentiation between intramolecular and intermolecular NOEs, a second NOESY experiment was measured without decoupling on ^{13}C during ^1H evolution, yielding doublets for protein-protein NOEs but single peaks for detergent-protein NOEs. Arg^{43} for this measurement is shown in comparison with Arg^{43} in (B), showing the presence of a protein-protein NOE at 0.8 ppm and the presence of a detergent-protein NOE at 1.2 ppm.

nels, receptor serine kinases of the transforming growth factor–β (TGF-β) superfamily, and G protein–coupled receptors (GPCRs) (Fig. 5A). Although expression success varied according to induction conditions, proteolytic susceptibility of the target gene, and the length of the amino acid linker from Mistic to the fusion protein, in most cases (15 of 22 tested constructs, table S2) the desired product could be isolated from the membrane fraction of recombinant bacteria at yields exceeding 1 mg per liter of culture (Fig. 5B). The *Aplysia* potassium channel, aKv1.1, was extracted and purified in LDAO to verify that the expressed proteins resemble their native conformations; size exclusion chromatography (Fig. 5C) showed that it retains a tetrameric assembly. Additionally, several TGF-β receptors were found to retain native ligand-binding affinity and specificity (*22*). Taken in combination with the fact that all of these proteins partition to the membrane fraction of cell extracts, we conclude that there exists a high propensity for this system to produce IM proteins fully folded in their native conformations.

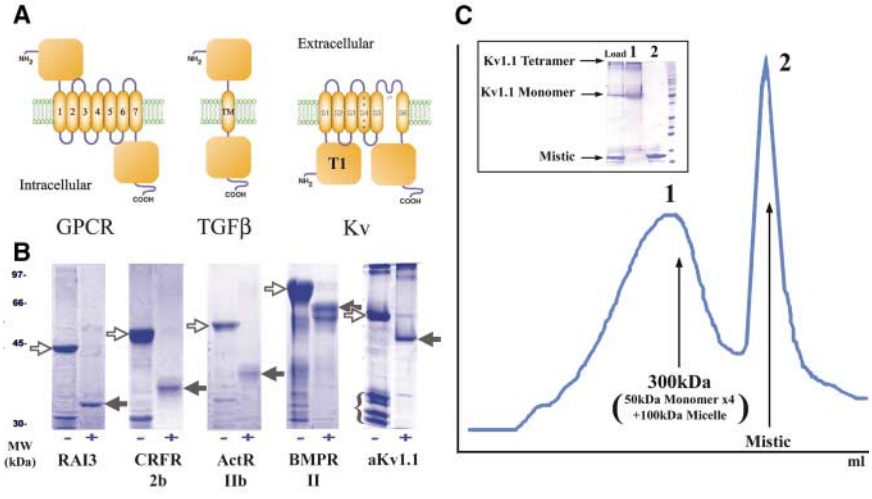


Fig. 5. Mistic-assisted eukaryotic IM protein expression. (**A**) Topological depictions of the three protein classes studied in this report: GPCRs, TGF-β family receptors, and voltage-gated K^+ channels (Kv). (**B**) SDS-PAGE results for various eukaryotic IM proteins. Lane pairs reveal expression of the desired protein from LDAO-solubilized membrane fractions after purification by Ni-NTA affinity chromatography. The Mistic-fused protein is shown on the left (open arrow); the final product after removal of Mistic by thrombin digestion is on the right (solid arrow). Protein identities were verified for select samples [including retinoic acid–induced protein 3 (RAI3), bone morphogenetic protein receptor type II (BMPR II), and aKv1.1] by N-terminal Edman degradation sequencing of at least 14 residues of the target protein after separation from Mistic. The additional bands in the sample of aKv1.1 before digestion (bracket) were determined to be truncated products containing fragments of the N-terminal T1 domain of this channel. The region between T1 and the membrane-spanning domains of this channel is known to be flexible and proteolytically susceptible. (**C**) Gel filtration profile of thrombin-digested aKv1.1 run in 3 mM LDAO on a Superose-6 column. aKv1.1 elutes as a detergent solubilized tetramer subsequent to Mistic removal. (Inset) Baseline separation between aKv1.1 (lane 1) and Mistic (lane 2) allows two-step purification of aKv1.1 to near-homogeneity.

To validate Mistic's direct role in assisting in the production of these recombinant IM proteins, we introduced mutations at three potentially structurally disruptive sites within the core of the protein (Figs. 2A and 6A). Expression tests of these Mistic variants, alone and fused to aKv1.1, indicate that the integrity of Mistic's structure is essential to its ability to chaperone cargo proteins to the bacterial lipid bilayer (Fig. 6B). The single mutation of a core methionine (Met^{75}) to alanine, in particular, sufficiently destabilized Mistic's structure such that it partitioned between the membrane and the cytoplasm. This same mutant yielded no protein expression when fused to aKv1.1, confirming that Mistic's structure and resulting membrane affinity are critical for its ability to facilitate the production of target IM proteins.

Given the highly acidic surface of Mistic (Fig. 3, B and C), it is still conceivable that the conformation of Mistic in the cell membrane differs from the structure observed in the Mistic-detergent complex. Recently, charged transmembrane helices have been shown to play dynamic roles within the lipid bilayer in ion channels and transporters (*23*, *24*). Conformational flexibility, such as rotation of the four helices about their helical axes or even partial unraveling of the helical bundle, may allow Mistic to adapt

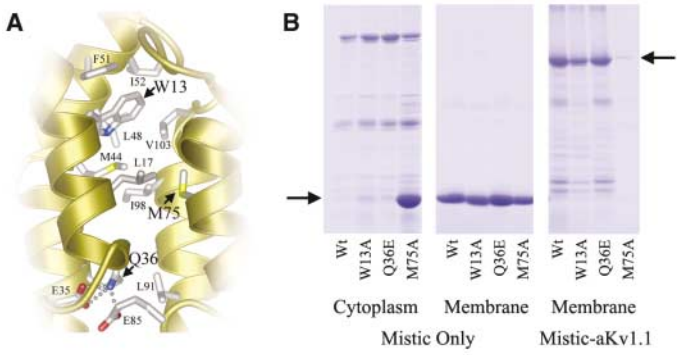


Fig. 6. Mutational disruption of Mistic's structure and function. (**A**) Residues forming the core of Mistic, with those mutated in structural disruption studies highlighted with arrows. (**B**) Mistic mutated singly at three core residues displays varying structural stability and functionality. Mutation of Trp^{13} to Ala (W13A) reduces the overall yield of fused aKv1.1 by a factor of 2 to 3. More important, mutation of Met^{75} to Ala (M75A) destabilizes the structure of Mistic sufficiently such that, when expressed by itself, it partitions substantially into the cytoplasm (fourth lane from left), in stark contrast to wild-type Mistic or any of the other mutants analyzed. This results in a functionally disabled protein; thus, when M75A is fused to aKv1.1, there is no detectable yield of this protein (rightmost lane).

to the lipid environment in a fashion analogous to the mechanisms of membrane integration for the chloride channel CLIC1 (*25*) or diphtheria toxin (*26*), both of which exist alternately in soluble and membrane-integrated forms. Molecular interplay between lipid composition and membrane insertion of IM protein structures is another intriguing possibility (*27*). Although complete understanding of the integration dynamics of Mistic requires further study, all available data suggest that it must autonomously associate with the bacterial membrane and that this property alone accounts for its high efficiency in chaperoning the production and integration of downstream cargo proteins (fig. S6). Taken together with the NMR techniques and protocols developed and used for Mistic structure determination, Mistic's unique ability to assist in the production of IM proteins opens new avenues around traditional obstacles in the study of IM proteins, particularly those of eukaryotic origin.

References and Notes

1. C. G. Tate, *FEBS Lett.* **504**, 94 (2001).
2. R. Laage, D. Langosch, *Traffic* **2**, 99 (2001).
3. H. Kiefer, R. Vogel, K. Maier, *Receptors Channels* **7**, 109 (2000).
4. J. Tucker, R. Grisshammer, *Biochem. J.* **317**, 891 (1996).
5. W. C. Wimley, S. H. White, *Biochemistry* **39**, 4432 (2000).
6. K. Wuthrich, *Nature Struct. Biol.* **5**, 492 (1998).
7. G. M. Clore, A. M. Gronenborn, *Nature Struct. Biol.* **4**, 849 (1997).
8. V. K. Rastogi, M. E. Girvin, *Nature* **402**, 263 (1999).
9. K. R. MacKenzie, J. H. Prestegard, D. M. Engelman, *Science* **276**, 131 (1997).
10. C. Fernandez, C. Hilty, G. Wider, P. Guntert, K. Wuthrich, *J. Mol. Biol.* **336**, 1211 (2004).
11. P. M. Hwang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 13560 (2002).
12. A. Arora, F. Abildgaard, J. H. Bushweller, L. K. Tamm, *Nature Struct. Biol.* **8**, 334 (2001).
13. See supporting data on *Science* Online.
14. V. Ramamurthy, D. Oliver, *J. Biol. Chem.* **272**, 23239 (1997).
15. K. Pervushin, R. Riek, G. Wider, K. Wuthrich, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 12366 (1997).
16. M. Salzmann, G. Wider, K. Pervushin, K. Wuthrich, *J. Biomol. NMR* **15**, 181 (1999).
17. C. Ritter, T. Luhrs, W. Kwiatkowski, R. Riek, *J. Biomol. NMR* **28**, 289 (2004).
18. P. A. Kosen, *Methods Enzymol.* **177**, 86 (1989).
19. J. L. Battiste, G. Wagner, *Biochemistry* **39**, 5355 (2000).
20. P. Guntert, *Methods Mol. Biol.* **278**, 353 (2004).
21. C. Hilty, G. Wider, C. Fernandez, K. Wuthrich, *Chembiochem* **5**, 467 (2004).
22. T. P. Roosild *et al.*, data not shown.
23. R. B. Bass, P. Strop, M. Barclay, D. C. Rees, *Science* **298**, 1582 (2002).
24. Y. Jiang *et al.*, *Nature* **423**, 33 (2003).
25. D. R. Littler *et al.*, *J. Biol. Chem.* **279**, 9298 (2004).
26. S. Choe *et al.*, *Nature* **357**, 216 (1992).
27. W. Zhang, M. Bogdanov, J. Pi, J. Pittard, W. Dowhan, *J. Biol. Chem.* **278**, 50128 (2003).

28. We thank E. Wiater for help in TGF-β receptor binding studies and C. Park for Edman degradation sequencing of proteins. Supported by NIH grant GM056653. R.R. is a Pew scholar. The bundle of 10 conformers representing the NMR structure is deposited in the PDB database with accession code 1YGM. The coding sequence of Mistic has been deposited in GenBank with accession code AY874162.

Supporting Online Material

www.sciencemag.org/cgi/content/full/307/5713/1317/DC1
Materials and Methods
Figs. S1 to S6
Tables S1 and S2
References

14 October 2004; accepted 6 January 2005
10.1126/science.1106392

The Genome of the Basidiomycetous Yeast and Human Pathogen *Cryptococcus neoformans*

Brendan J. Loftus,[1]* Eula Fung,[2] Paola Roncaglia,[3] Don Rowley,[2] Paolo Amedeo,[1] Dan Bruno,[2] Jessica Vamathevan,[1] Molly Miranda,[2] Iain J. Anderson,[1] James A. Fraser,[4] Jonathan E. Allen,[1] Ian E. Bosdet,[5] Michael R. Brent,[6] Readman Chiu,[5] Tamara L. Doering,[7] Maureen J. Donlin,[8] Cletus A. D'Souza,[9] Deborah S. Fox,[4,10] Viktoriya Grinberg,[1] Jianmin Fu,[11] Marilyn Fukushima,[2] Brian J. Haas,[1] James C. Huang,[4] Guilhem Janbon,[12] Steven J. M. Jones,[5] Hean L. Koo,[1] Martin I. Krzywinski,[5] June K. Kwon-Chung,[13] Klaus B. Lengeler,[4,14] Rama Maiti,[1] Marco A. Marra,[5] Robert E. Marra,[4,15] Carrie A. Mathewson,[5] Thomas G. Mitchell,[4] Mihaela Pertea,[1] Florenta R. Riggs,[1] Steven L. Salzberg,[1] Jacqueline E. Schein,[5] Alla Shvartsbeyn,[1] Heesun Shin,[5] Martin Shumway,[1] Charles A. Specht,[16] Bernard B. Suh,[17] Aaron Tenney,[6] Terry R. Utterback,[18] Brian L. Wickes,[11] Jennifer R. Wortman,[1] Natasja H. Wye,[5] James W. Kronstad,[9] Jennifer K. Lodge,[8] Joseph Heitman,[4] Ronald W. Davis,[2] Claire M. Fraser,[1] Richard W. Hyman[2]

Cryptococcus neoformans is a basidiomycetous yeast ubiquitous in the environment, a model for fungal pathogenesis, and an opportunistic human pathogen of global importance. We have sequenced its ~20-megabase genome, which contains ~6500 intron-rich gene structures and encodes a transcriptome abundant in alternatively spliced and antisense messages. The genome is rich in transposons, many of which cluster at candidate centromeric regions. The presence of these transposons may drive karyotype instability and phenotypic variation. *C. neoformans* encodes unique genes that may contribute to its unusual virulence properties, and comparison of two phenotypically distinct strains reveals variation in gene content in addition to sequence polymorphisms between the genomes.

With an increased immunocompromised population as a result of AIDS and widespread immunosuppressive therapy, *Cryptococcus neoformans* has emerged as a major pathogenic microbe in patients with impaired immunity (*1*). *C. neoformans* elaborates two

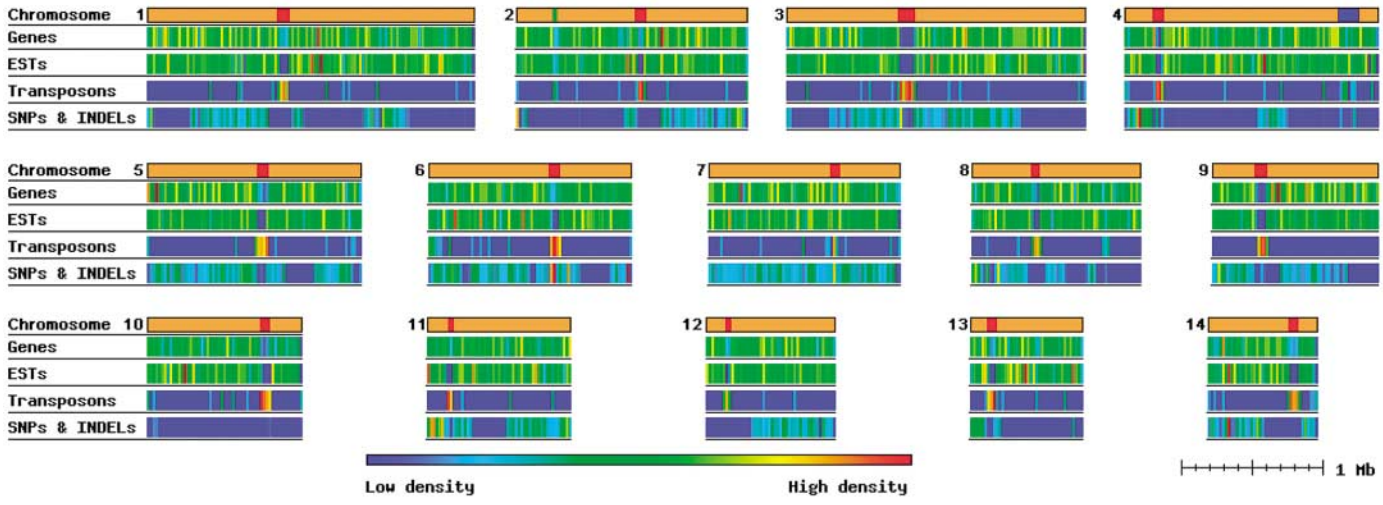


Fig. 1. The *C. neoformans* JEC21 genome with each chromosome represented as a colored bar. Specific features are pseudocolored, from red (high density) to deep blue (low density) and plotted on a log scale. These include the density of genes, transposons, expressed sequence tags (ESTs), and predicted SNPs and indels. Candidate centromeric regions and the MAT locus are represented as red bars and a blue bar, respectively. The location of the rDNA repeat is represented by a green bar.

specialized virulence factors, a polysaccharide capsule (*2*) and the antioxidant pigment melanin (*3*), which enhance human infection and central nervous system colonization. Here, we report the genome sequence of two related strains of *C. neoformans* serotype D (JEC21 and B-3501A) as an important step in the elucidation of the genomic basis for virulence in this pathogenic yeast.

The 19-Mb genome sequence of *C. neoformans* JEC21 [excluding the ribosomal RNA (rDNA) repeats region constituting ~5% of the genome] spans 14 chromosomes from 762 kb to 2.3 Mb (table S1), whereas the 18.5-Mb sequence of the B-3501A strain consists of 14 linked assemblies (scaffolds). Unlike *S. cerevisiae*, the genome of *C. neoformans* shows no evidence for a whole-genome duplication (*4*). However, a chromosomal translocation and an exact ~60-kb segmental duplication are present in JEC21 compared with B-3501A (*5*). Almost 5% of the genome consists of transposons, the majority clustered on each chromosome in single blocks that span 40 to 100 kb that may represent sequence-independent regional centromeres, similar to those in *S. pombe* and *N. crassa* (*6*) (Fig. 1). Each block is unique but all contain at least one copy of the Tcn5 or Tcn6 transposons, which may represent functional elements or target the centromeres. Transposons are also clustered adjacent to the rDNA repeats and within the mating-type (MAT) locus (Fig. 1). In contrast to the other transposons, the long interspersed nuclear element–like (LINE-like) retroelement Cnl1 shows a marked preference for telomeric regions.

To ensure accurate gene structure annotation, sequence data were obtained from both ends of more than 23,000 cDNA clones of a full-length normalized cDNA library from *C. neoformans* JEC21 cells grown under various conditions (*7*). A total of 6572 protein-encoding genes were identified, which contain an average of 6.3 exons of 255 base pairs (bp) and 5.3 introns of 67 bp (table S2). The mean transcript size of 1.9 kb contains an average of 15% noncoding sequence from both the 5′ and 3′ ends. The gene organization in *C. neoformans* is thus considerably more complex than that of ascomycetes for which genome sequence (table S2) is available and is comparable to that observed in *Arabadopsis thaliana* or *Caenorhabditis elegans*.

A conspicuous feature to emerge from comparing cDNA and genome sequence data is evidence for alternative splicing and endogenous antisense transcripts, in some cases emanating from the same gene locus (Fig. 2). Alternative splicing and natural antisense RNA transcribed in cis were identified in genes encoding diverse functions distributed genome-wide, which suggests that both are widespread genetic regulatory mechanisms in *C. neoformans* (tables S3 to S5). Alternative splice forms were predicted for 277 genes, or 4.2% of the transcriptome (table S4), and a variety of mechanisms could be identified (e.g., exon skipping, truncation, and extension at both 5′ and 3′ ends). Antisense transcripts were identified for 53 genes; however, they appear to have no appreciable coding potential and are usually completely overlapped by their sense counterparts (table S5). The presence and frequency of these antisense transcripts and the presence of the molecular components necessary for RNA interference extend previous studies (*8*) and indicate that regulation by double-stranded RNA is likely a general regulatory mechanism in this organism.

JEC21 and B-3501A are highly related inbred strains of the alpha mating type, the most prevalent mating type in environmental

[1]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [2]Stanford Genome Technology Center, Stanford University, 855 California Avenue, Palo Alto, CA 94304, USA. [3]Neurobiology Sector, International School for Advanced Studies (SISSA-ISAS), Via Beirut 2-4, 34014 Trieste, Italy. [4]Department of Molecular Genetics and Microbiology, Duke University Medical Center, 322 CARL Building, Research Drive, Box 3546, DUMC, Durham, NC 27710, USA. [5]Genome Sciences Centre, 100-570 West 7th Avenue, Vancouver, BC V5Z 4S6, Canada. [6]Laboratory for Computational Genomics, Washington University, One Brookings Drive, St. Louis, MO 63130, USA. [7]Department of Molecular Microbiology, Washington University School of Medicine, 660 South Euclid Avenue, St. Louis, MO 63110, USA. [8]Department of Biochemistry and Molecular Biology, Saint Louis University School of Medicine, 1402 S. Grand Boulevard, St. Louis, MO 63104, USA. [9]The Michael Smith Laboratories, The University of British Columbia, 2185 East Mall, Vancouver, BC V6T 1Z4, Canada. [10]Research Institute for Children and the Department of Pediatrics, Louisiana State Health Science Center, Children's Hospital, 200 Henry Clay Avenue, New Orleans, LA 70118, USA. [11]University of Texas Health Science Center, 7703 Floyd Curl Drive, San Antonio, TX 78229, USA. [12]Unité de Mycologie Moléculaire, Institut Pasteur, 25 rue du Docteur Roux, Cedex 15, Paris, France. [13]Molecular Microbiology Section, Laboratory of Clinical Investigation, National Institutes of Health (NIAID/NIH), 9000 Rockville Pike, Bethesda, MD 20892, USA. [14]Institut für Mikrobiologie, Heinrich-Heine-Universität, Universitätsstraße 1/26.12, Düsseldorf, Germany. [15]Plant Pathology and Ecology, The Connecticut Agricultural Experiment Station, 123 Huntington Street, New Haven, CT 06511, USA. [16]Department of Medicine, Boston University, 650 Albany Street, EBRC-625, Boston, MA 02118, USA. [17]Department of Biomolecular Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064 USA. [18]Joint Technology Center, J. Craig Venter Foundation, 5 Research Place, Rockville, MD 20850, USA.

*To whom correspondence should be addressed. E-mail: bjloftus@tigr.org

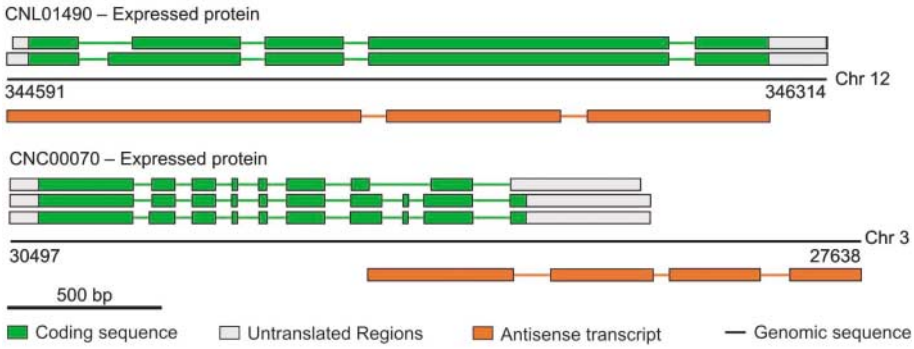


Fig. 2. Gene structures that display evidence for both alternative splicing and natural in cis antisense transcripts based on JEC21 cDNA alignments to the genome sequence. Colored boxes represent exonic regions. Each gene structure represents an alternative spliced form. The black line represents the genomic sequence.

and clinical isolates (*9*). As a result of backcrossing during strain construction, the sequence differences that distinguish these strains are restricted to 50% of their genomes, which overall are 99.5% identical at the sequence level. The predicted single-nucleotide polymorphisms (SNPs) and insertion and deletion polymorphisms (indels) are distributed in blocks of high and low sequence polymorphism, reflecting the recombination events that occurred during production of these sibling strains (Fig. 1). The phenotypes of JEC21 and B-3501 differ markedly, with B-3501A being more thermotolerant and more virulent in animal models than JEC21. To investigate the genetic basis for these differences, genomic regions encompassing JEC21 genes were compared directly with the B-3501A assembly. The vast majority (99.7%) of genes share >98% nucleotide identity (fig. S1). Strain-specific genes were experimentally verified by polymerase chain reaction and included a Ras guanosine triphosphatase–activating protein and two proteins of unknown function specific to B-3501A, whereas four proteins of unknown function were specific to JEC21. These genes, in addition to 22 duplicated genes in JEC21 located on the ~60-kb segmental duplication, delineate the strains.

A remarkable feature of *C. neoformans* is the link between virulence and mating type, which is governed by a specialized genomic region, the MAT locus (*10*). Genome analysis revealed several additional genes in MAT. Numerous other genes involved in mating are not in MAT or on the MAT chromosome and are scattered throughout the genome. Consistent with classification as a heterothallic fungus that does not switch mating type, there are no silent mating-type cassettes.

The major virulence factor of *C. neoformans* is its extensive polysaccharide capsule, an elaborate and dynamic structure that surrounds the fungal cell wall that is unique among fungi that affect humans (*2*). Genome analysis identified more than 30 new genes likely involved in capsule biosynthesis, including a family containing seven members of the capsule-associated (*CAP64*) gene. The *CAP64* family appears to be restricted to basidiomycetes, and two members encode alternatively spliced forms (table S5). A second family of six capsule-associated (*CAP10*) genes appears restricted to a subset of fungi and is absent from other yeasts.

The cell wall is an essential and unique component of fungi, and most of the genes involved in the biosynthesis of cell-wall polysaccharides are conserved between the ascomycetes and *C. neoformans*, making them attractive targets for broad-spectrum antifungal drugs. However, *S. cerevisiae* and *C. neoformans* manifest notable differences in their mechanisms of cell-wall protein association. In *S. cerevisiae*, two major classes of proteins are covalently bound to the cell wall: the Pir proteins and a set of proteins that are covalently attached to the cell wall by a glycosylphosphatidylinositol (GPI) anchor. *C. neoformans* lacks both Pir-related genes and several genes that have been implicated in attachment of the GPI anchors to the β-1,6-glucan in the cell wall (*11*). Genome analysis also predicts more than 50 extracellular mannoproteins that may be associated with the cell wall, most of which are unique to *C. neoformans*.

The phylum Basidiomycota last shared a common ancestor with the ascomycetes ~900 million years ago, and the two phyla have diverged considerably (*12*). Overall, 65% of *C. neoformans* genes have conserved sequence homologs in a sampling of completed fungal genomes (table S2), and of these 12% are restricted to the basidiomycete genome *Phanerochaete chrysosporium*. Another 10% appear to be unique to *C. neoformans*, based on the absence of identifiable homologs in the current public databases, whereas the remaining 25% match nonfungal sequences (*7*). Lineage-specific gene family expansions do not represent the most abundant protein domains within the *C. neoformans* genome, which are similar to those of ascomycetous fungi (tables S6 and S7). Two of the 11 gene families that appear unique to *C. neoformans* are involved in capsule formation, and another encodes nucleotide sugar epimerases associated with cell-wall formation. About 60% of the *C. neoformans* genes could be assigned gene ontology terms for molecular function (*7*), and comparison with *S. cerevisiae* reveals a similar distribution of genes across nearly all functional categories (fig. S2). One exception is an expansion of the drug-efflux transporters of the major facilitator superfamily in *C. neoformans*, which suggests enhanced transport capability in this environmental yeast.

Recently, the *Candida albicans* genome was reported (*13*), enabling a comparison between these divergent pathogenic fungi. *C. neoformans* is an environmental organism that infects through inhalation, whereas *C. albicans* is part of normal human microbiota and infects by bloodstream invasion. Myriad cell-surface proteins implicated in *C. albicans* adhesion to epithelial cells are absent in *C. neoformans*, which suggests that *C. neoformans* binds host cells by distinct mechanisms. *C. neoformans* elaborates both capsule and melanin; *C. albicans* makes neither and lacks genes for their production.

The *C. neoformans* genome sequence provides new insights into this important fungal human pathogen. The genome encodes a core complement of genes common to other fungi and, despite a large divergence time, the functional distribution of many *C. neoformans* genes mirrors that of *S. cerevisiae*. By contrast with *S. cerevisiae*, however, the *C. neoformans* genome displays an intron-rich gene tapestry and a transcriptome rife with alternative splicing and antisense transcripts. These genome sequence data, together with those from another basidiomycete, *P. chrysosporium* (*14*), suggest that more complex gene structures may be a general feature of basidiomycetes (table S2). The genome sequence data described herein from two closely related strains of *C. neoformans* provide a foundation to explore the molecular basis of virulence in this pathogen and reveal differences in virulence strategies between *C. neoformans* and other pathogenic fungi.

References and Notes

1. A. Casadevall, J. R. Perfect, *Cryptococcus neoformans* (ASM Press, Washington, DC, 1998).
2. I. Bose, A. J. Reese, J. J. Ory, G. Janbon, T. L. Doering, *Eukaryot. Cell* **2**, 655 (2003).
3. A. Casadevall, A. L. Rosas, J. D. Nosanchuk, *Curr. Opin. Microbiol.* **3**, 354 (2000).
4. M. Kellis, B. W. Birren, E. S. Lander, *Nature* **428**, 617 (2004).
5. J. A. Fraser *et al.*, in preparation.
6. E. B. Cambareri, R. Aisner, J. Carbon, *Mol. Cell. Biol.* **18**, 5465 (1998).
7. Materials and methods are available as supporting material on *Science* Online.

8. J. M. Gorlach, H. C. McDade, J. R. Perfect, G. M. Cox, *Microbiol.* **148**, 213 (2002).
9. K. J. Kwon-Chung, J. E. Bennett, *Am. J. Epidemiol.* **108**, 337 (1978).
10. K. B. Lengeler *et al.*, *Eukaryot. Cell* **1**, 704 (2002).
11. S. Shahinian, H. Bussey, *Mol. Microbiol.* **35**, 477 (2000).
12. S. B. Hedges, J. E. Blair, M. L. Venturi, J. L. Shoe, *BMC Evol. Biol.* **4**, 2 (2004).
13. T. Jones *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 7329 (2004).
14. D. Martinez *et al.*, *Nature Biotechnol.* **22**, 695 (2004).
15. We thank J. Perfect, F. Dietrich, and J. Murphy for their invaluable and ongoing support for the *C. neoformans* genome project. Funding was provided by National Institute of Allergy and Infectious Diseases (NIAID) cooperative agreements AI48594 (C.M.F.) and AI47087 (R.W.D.). Accession numbers for the JEC21 genome (AE017341-AE017353, AE017356), the B-3501A genome (AAEY00000000), and the JEC21 cDNA sequences (CF675703.1-CF722528.1) have been submitted to GenBank.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1103773/DC1
Materials and Methods
Figs. S1 to S3
Tables S1 to S9
References

9 August 2004; accepted 5 January 2005
Published online 13 January 2005;
10.1126/science.1103773
Include this information when citing this paper.

Control of Excitatory and Inhibitory Synapse Formation by Neuroligins

Ben Chih, Holly Engelman, Peter Scheiffele*

The normal function of neural networks depends on a delicate balance between excitatory and inhibitory synaptic inputs. Synapse formation is thought to be regulated by bidirectional signaling between pre- and postsynaptic cells. We demonstrate that members of the Neuroligin family promote postsynaptic differentiation in cultured rat hippocampal neurons. Down-regulation of neuroligin isoform expression by RNA interference results in a loss of excitatory and inhibitory synapses. Electrophysiological analysis revealed a predominant reduction of inhibitory synaptic function. Thus, neuroligins control the formation and functional balance of excitatory and inhibitory synapses in hippocampal neurons.

Adhesion molecules bridge the pre- and postsynaptic compartments of synapses in the central nervous system. Neuroligin-1 (NL-1), a member of the Neuroligin family of postsynaptic adhesion molecules, can trigger formation of functional presynaptic terminals in axons through interaction with its axonal receptor β-neurexin [(*1*–*3*), reviewed in (*4*–*6*)]. To explore whether the β-neurexin–neuroligin complex acts bidirectionally and controls postsynaptic differentiation, we overexpressed NL-1 in cultured hippocampal neurons (*7*). Analysis of dendritic morphology, postsynaptic scaffolding molecules, and postsynaptic glutamate receptor distribution revealed that NL-1 promotes assembly of the postsynaptic apparatus (Fig. 1). NL-1–overexpressing neurons showed a 68 ± 7% increase in the density of dendritic spine–like protrusions. Spines in NL-1–expressing cells frequently exhibited irregular, hand-shaped heads with multiple presynaptic terminals labeled for the vesicular glutamate transporter 1 (vGlut1), a marker of excitatory synapses (Fig. 1A; fig. S1). The density of synaptic puncta containing the scaffolding proteins PSD-95 and Homer was increased significantly (Fig. 1B). Moreover, staining for the NR1 subunit of *N*-methyl-D-aspartate (NMDA) receptors revealed that NL-1 strongly promotes NMDA receptor recruitment (Fig. 1D). We also observed recruitment of AMPA-type glutamate receptors, as indicated by clustering of GluR2/3 subunits in some NL-1–expressing cells. However, high NL-1 levels

Department of Physiology and Cellular Biophysics, Center for Neurobiology and Behavior, Columbia University, New York, NY 10032, USA.

*To whom correspondence should be addressed. E-mail: ps2018@columbia.edu

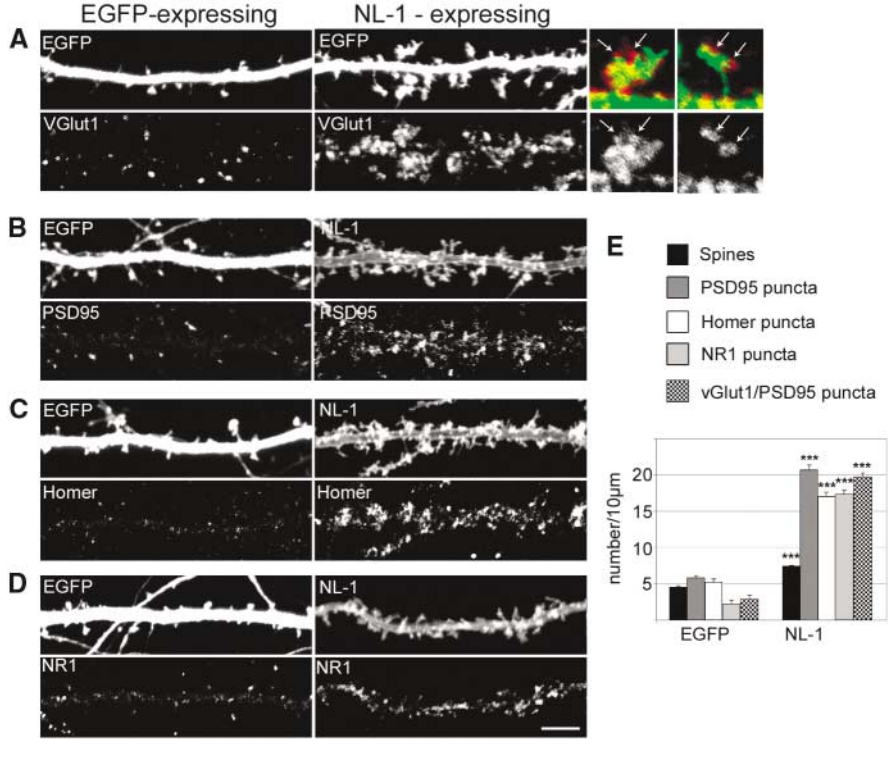


Fig. 1. NL-1 promotes postsynaptic differentiation. Hippocampal neurons were cotransfected with expression vectors for hemagglutinin (HA)–tagged NL-1 and EGFP or with EGFP vectors only. (**A**) Immunostaining for vGlut1 and EGFP in control cells expressing EGFP (left column) and cells coexpressing EGFP and NL-1 (right column). NL-1–induced spine structures contacting multiple presynaptic terminals (right). (**B**) Immunostaining for PSD-95 and EGFP (left) or HA epitope to detect NL-1 (right). (**C**) Immunostaining for Homer and EGFP (left) or HA epitope to detect NL-1 (right). (**D**) Immunostaining for NMDA-receptor subunit 1 (NR1) and EGFP (left) or HA epitope to detect NL-1 (right). Scale bar, 5 μm. (**E**) Quantification of postsynaptic protein recruitment, dendritic spine induction, and synapse formation in cells expressing NL-1 and EGFP-transfected control cells. vGlut1/PSD-95 shows density of puncta with colocalizing pre- and postsynaptic markers (SEM, $n = 10$, $^{***}P < 0.001$).

led to dispersion of GluR2/3, likely due to the depletion of cytoplasmic binding partners for NL-1 (*8*). In summary, these experiments show that NL-1 is a potent inducer of excitatory postsynaptic differentiation.

NL-1–induced postsynaptic differentiation may be entirely mediated through scaffolding proteins interacting with the intracellular tail of NL-1 or may require extracellular interactions with the β-neurexin–NL-1 complex. To investigate the respective contributions of the NL-1 extracellular and intracellular domains to postsynaptic differentiation, we analyzed two NL-1 mutants. In the mutant NL-swap, the extracellular cholinesterase domain of NL-1 was exchanged with the homologous sequence from acetylcholinesterase to yield a mutant NL protein in which the β-neurexin–binding site was inactivated (*1*). In the intracellular mutant NLΔC, the cytoplasmic tail of NL-1 was truncated, which removed the PDZ-binding motif that interacts with postsynaptic scaffolding molecules such as PSD-95 (*9*). When overexpressed in hippocampal neurons, the extracellular mutant NL-swap could still stimulate the recruitment of PSD-95 into clusters at the cell membrane. However, these PSD-95 clusters were largely extrasynaptic and did not align with the presynaptic marker vGlut1 (Fig. 2A). This result suggests that this NL swap acts as a dominant-negative mutant that uncouples nucleation of the postsynaptic scaffold from the presynaptic terminal. Expression of the intracellular mutant NLΔC did not stimulate PSD-95 clustering, which confirmed that the cytoplasmic tail of NL-1 is essential for PSD-95 recruitment (Fig. 2A). Despite the inability to recruit PSD-95, NLΔC increased NMDA-receptor cluster density at synapses, albeit less efficiently than did the wild-type protein (Fig. 2, B and C). These findings suggest that NMDA receptors are primarily recruited to NL-1–induced synapses independently of PSD-95 and that synaptic recruitment requires the ability of NL-1 to interact with β-neurexin or additional extracellular ligands. Therefore, the β-neurexin–neuroligin complex provides a nucleation site for the assembly of postsynaptic scaffolding molecules and NMDA receptors opposite the presynaptic terminal through intracellular and extracellular interactions.

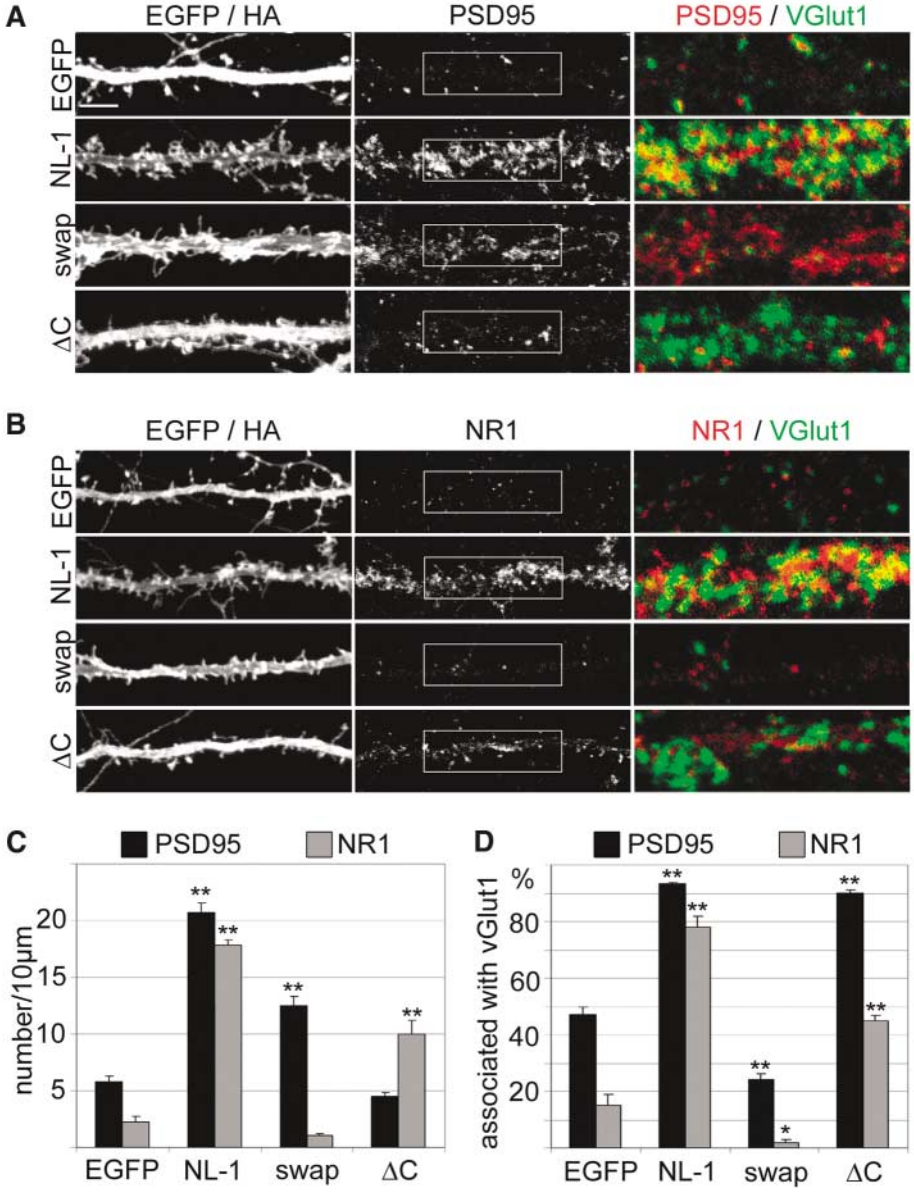


Fig. 2. Perturbation of postsynaptic assembly by NL-1 mutants. Hippocampal neurons expressing EGFP, HA-NL-1 [wild type (wt)], extracellular mutant (swap), or a C-terminally deleted mutant (ΔC). (**A**) Immunostaining for vGlut1, PSD-95, and EGFP or the HA epitope. Boxed area was enlarged in third column. Scale bar, 5 μm. (**B**) Immunostaining for vGlut1, NR1, and EGFP or the HA epitope. (**C** and **D**) Quantification of clustering (C) and synaptic localization (D) of PSD-95 and NR1 in cells expressing wild-type or mutant NL as compared with EGFP-expressing control cells (SEM, $n = 10$, $^{**}P < 0.01$).

To evaluate the consequences of reduced neuroligin function, we used RNA interference (*10*, *11*). We generated small-hairpin RNAs (shRNAs) directed against the rodent neuroligin isoforms NL-1, NL-2, and NL-3 and tested their efficiency and specificity by cotransfection with neuroligin expression vectors into HEK293 cells (Fig. 3A). Knockdown of neuroligin expression was strictly isoform-specific, e.g., NL-1–directed shRNAs did not alter NL-2 and NL-3 expression. Introduction of the shRNAs into hippocampal neurons in culture confirmed the suppression of endogenous neuroligins individually and in combination (Fig. 3B; fig. S2). Expression of other neuronal proteins such as PSD-95 or class III β-tubulin was not altered, and introduction of the shRNA vector lacking an insert or containing a control shRNA had no effect on any of the proteins analyzed.

Suppression of single or multiple neuroligin isoforms reduced excitatory synapse formation. When the shRNAs were transfected into hippocampal neurons, we observed fewer vGlut1-positive excitatory presynaptic terminals (Fig. 3, C and D). Introduction of an expression vector lacking the shRNA insert or targeting an unrelated mRNA did not alter the density of vGlut1 puncta. Knockdown of each of the neuroligin isoforms also inhibited postsynaptic maturation, as indicated by a significant reduction in the density of dendritic spines, although this was less severe for suppression of NL-3. Simultaneous knockdown of all three rodent neuroligin isoforms resulted in a 70% reduction in the number of morphologically recognizable excitatory synapses as detected by colocalization of vGlut1 and the glutamate receptor subunit GluR1 (Fig. 3E). The number of synapses in triple-neuroligin-knockdown cells was also

reduced when cultures were maintained in tetrodotoxin (TTX) to block all sodium channel–dependent action potentials (Fig. 3E). Synapse loss is therefore not a secondary consequence of an essential function of neuroligins in action potential–dependent neurotransmission.

It was interesting that the reduction of excitatory terminal density in triple-knockdown cells was not more severe than in the single-knockdown cells (Fig. 3, C and D). This indicated that the function of all three neuroligin isoforms might be coupled or that a critical level of total neuroligin proteins might be required for normal function. To further investigate this, we tested whether the triple-neuroligin-knockdown phenotype could be rescued by cotransfection of a human NL-3 cDNA, which is resistant to sh-NL3–directed cleavage. Simultaneous transfection of human NL-3 with shRNAs against NL-1, -2, and -3 restored dendritic spine density and vGlut1 clustering, whereas cotransfection of an inactive NL mutant (NL-swap) did not rescue the knockdown phenotype (Fig. 3, C and D). Similarly, defects caused by knockdown of only NL-3 could be suppressed by overexpression of NL-1 (*8*). Increasing expression of single neuroligin isoforms can, therefore, compensate for defects caused by the loss of other neuroligins. In summary, these experiments demonstrate that neuroligin isoforms are critical for normal excitatory synapse formation and have partially overlapping functions.

Recent studies suggested a potential role for neuroligins at inhibitory synapses (*12*, *13*). Immunostaining for endogenous NL-2 and the vesicular γ-aminobutyric acid transporter (VGAT) confirmed that the NL-2 isoform is concentrated at inhibitory synapses (Fig. 4A). By contrast, NL-1 is concentrated at mature excitatory synapses (*14*). When transfected into hippocampal neurons, all NL isoforms stimulated the formation of both excitatory and inhibitory terminals (Fig. 4B; fig. S3). NL-2 was more effective than NL-1 or NL-3 with respect to inhibitory terminal induction (Fig. 4, C and D) and might, therefore, preferentially contribute to inhibitory synapse formation. However, each of the three neuroligin isoforms is capable of inducing both excitatory and inhibitory terminals when expressed at a sufficiently high level.

Using shRNA-mediated knockdown revealed an essential function for neuroligins in inhibitory synapse formation. Suppression of any single neuroligin isoform resulted in a reduction in the density of VGAT-positive presynaptic terminals (Fig. 5, A and B). Among the individual isoforms, this effect was most significant for suppression of NL-2. Simultaneous knockdown of all three neuroligin isoforms resulted in the most robust reduction of inhibitory terminals. In a way similar to excitatory synapses, we found that some inhibitory terminals could form, even when all neuroligin isoforms were suppressed. We cannot exclude residual neuroligin levels (<10%) that escaped the shRNA-directed down-regulation as the cause, but we consider it more likely that there are other synapse-inducing proteins that promote

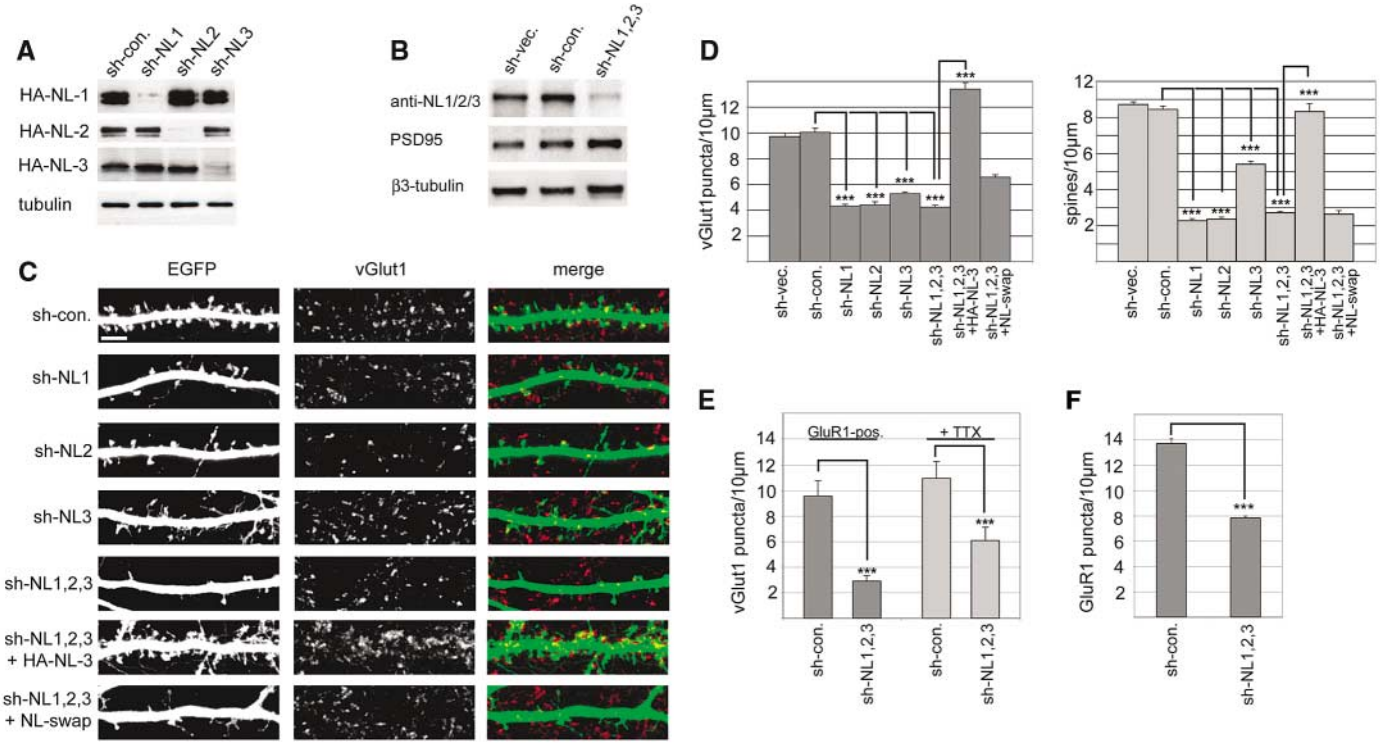


Fig. 3. Suppression of NL isoforms leads to excitatory synapse loss. (**A**) HEK293 cells cotransfected with HA-tagged NL-1, -2, or -3 and NL shRNAs (sh-NL1, sh-NL2, sh-NL3). Cell lysates were probed with antibodies against HA and tubulin. The control shRNA (sh-con.) targets p53. (**B**) Hippocampal neurons were triple-infected with lentiviruses encoding shRNAs that are targeting NL1, NL2, and NL3. Total protein levels of NL-1, -2, and -3, as well as PSD-95 and β3-tubulin, were analyzed by Western blotting. No suppression is observed with a shRNA vector without hairpin insert (sh-vec.) and a control shRNA (sh-con., targeting p53). This experiment underestimates the efficiency of NL protein suppression, because the viruses only infect 90% of all neurons in the culture. (**C**) Hippocampal neurons were transfected with shRNAs directed against individual NL isoforms or triple-transfected with shRNAs for all three NL isoforms (sh-NL1,2,3) and immunostained with antibodies to vGlut1. Loss of vGlut-1–positive puncta and dendritic spines can be rescued by cotransfection of HA-NL-3 (sh-NL1-1 + HA-NL-3) but not by cotransfection of the inactive mutant NL-swap (sh-NL1,2,3 + NL-swap). Scale bar, 5 µm. (**D**) Quantification of vGlut1 (left) and dendritic spine density (right) in hippocampal neurons with reduced NL expression (SEM, $n = 10$, $^{***}P < 0.001$). To avoid including cells with NL-3 overexpression, we focused on the 20% of all HA-NL-3–positive cells with lowest HA-NL-3 expression levels. (**E**) Quantification of vGlut1 puncta that are also positive for the postsynaptic AMPA-receptor subunit GluR1 (GluR1-pos.) in control (sh-con.) and triple-neuroligin-knockdown cells (sh-NL1,2,3) and the effect of triple-neuroligin knockdown on vGlut1 clustering in the presence of 2 µM TTX (SEM, $n = 10$, $^{***}P < 0.001$). (**F**) Total GluR1 puncta density in hippocampal neurons transfected with control shRNA (sh-con.) or shRNAs targeting NL-1,2,3 (sh-NL1,2,3) (SEM, $n = 10$, $^{***}P < 0.001$).

neuroligin-independent synapse formation (*15*). However, acute suppression of neuroligins does result in a substantial reduction of excitatory and inhibitory synapse numbers, demonstrating that they are important regulators of synaptogenesis.

Because the knockdown of neuroligins affected the numbers of morphologically recognizable excitatory and inhibitory terminals, we investigated whether synaptic transmission was altered. We recorded mIPSCs (miniature inhibitory postsynaptic currents) and mEPSCs (miniature excitatory postsynaptic currents) in triple-neuroligin-knockdown (sh-NL1,2,3) hippocampal neurons (Fig. 5C). In knockdown cells, mIPSC amplitudes and frequency were 52 ± 21% and 93 ± 30% reduced, respectively (Fig. 5D). In comparison, mEPSCs were only slightly affected (Fig. 5D). These selective functional defects resulted in a shift in the balance of excitatory and inhibitory events in the NL-1, -2, and -3 knockdown cells: mEPSCs represented 36.0 ± 3.1% of the total number of synaptic events in control cells and 83.0 ± 3.6% of events in the neuroligin-knockdown cells (Fig. 5D). Combined with the data described above, these results lead to three main conclusions: (i) Neuroligins are potent inducers of postsynaptic differentiation; (ii) The neuroligin isoforms support both excitatory and inhibitory synapse formation; and (iii) Loss of neuroligin isoforms alters the normal excitatory/inhibitory balance in hippocampal neurons.

NL-2 preferentially localizes to inhibitory synapses (*13*). Our overexpression and knockdown experiments revealed effects of NL-2 on both excitatory and inhibitory synapses, although effects on inhibitory synapses were more prominent. Most likely, initial interactions with excitatory and inhibitory axons are somewhat promiscuous, because mismatched excitatory and inhibitory pre- and postsynaptic components are observed frequently in developing neurons (*16*–*18*). The preference of individual neuroligin isoforms for excitatory or inhibitory synapses might then be reinforced during synaptic maturation.

Loss of neuroligins had a selective effect on inhibitory synapse function. Despite a 70% decrease in the density of morphologically recognizable excitatory synapses in triple-neuroligin-knockdown cells (Fig. 3E), we did not observe an equivalent reduction in mEPSC frequency. This suggests the preferential loss of inactive and/or silent excitatory synapses or that the remaining excitatory synapses may exhibit increased activity. In either case, the selective reduction of inhibitory synapse function resulted in a significant imbalance of excitatory and inhibitory transmission. This observation is notable, because the excitatory to inhibitory (E/I) input ratio is critical for normal computation of neuronal excitation (*19*, *20*) and is generally kept constant by a homeostatic feedback mechanism (*20*, *21*). A recent study reported a similar alteration of the E/I ratio in cells in which expression of PSD-95 had been suppressed by RNA interference (*12*). The selective decrease in inhibitory synapse function in neuroligin-knockdown cells may indicate that formation or stabilization of functional inhibitory synapses relies more heavily on neuroligins than does the function of excitatory synapses. An alternative hypothesis is that neuroligins contribute to the homeostatic mechanism that maintains the E/I balance.

Inactivating mutations in human neuroligins are associated with autism spectrum disorders (*22*–*25*), as are perturbations in the E/I ratio and morphological aberrations in dendritic spines (*26*–*28*). Our work links neuroligin function with these two phenotypes in hippocampal neurons. Further analysis of the cellular defects caused by reduced neuroligin expression in vitro may, therefore, provide a useful framework for understanding the cellular defects of autism spectrum disorders.

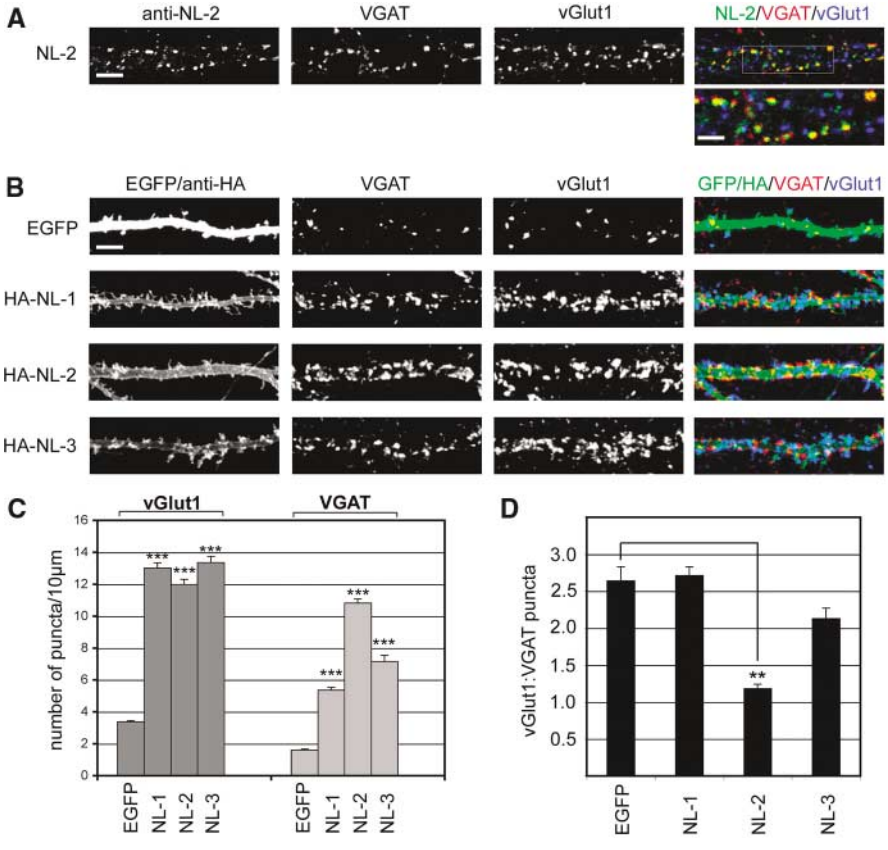
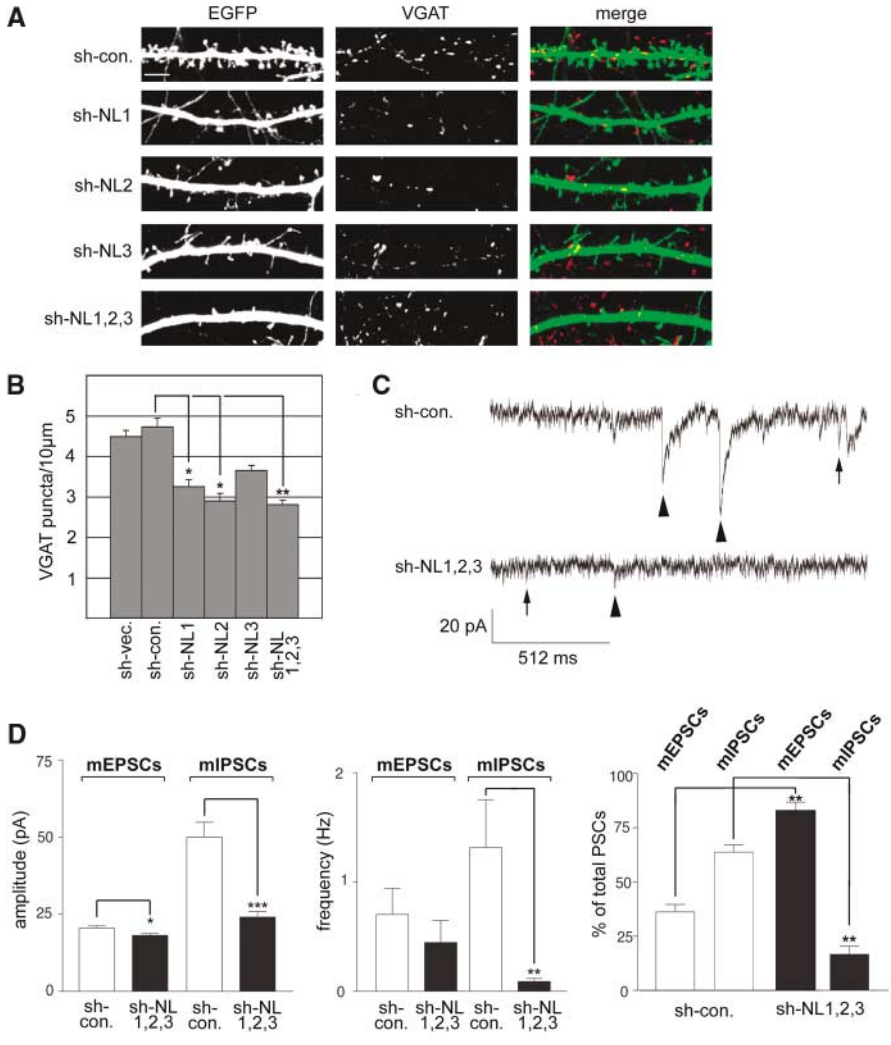


Fig. 4. NL-2 preferentially promotes inhibitory synapse formation. (**A**) Hippocampal neurons immunostained with antibodies to NL-2 (green), VGAT (red), and vGlut1 (blue). Right panel shows merge at higher magnification. Scale bar, 5 μm (left) and 10 μm (right). (**B**) Hippocampal neurons transfected with expression vectors for EGFP, HA-tagged NL-1, NL-2, or NL-3 were immunostained with antibodies to the HA epitope (green), VGAT (red), and vGlut1 (blue). Scale bar, 5 μm. (**C**) Density of excitatory and inhibitory terminals for cells overexpressing EGFP, NL-1, NL-2, or NL-3, (SEM, $n = 10$ cells, $^{***}P < 0.001$). (**D**) Ratio of vGlut1:VGAT–positive terminals in NL-overexpressing cells ($n = 10$ cells, $^{**}P < 0.01$).

References and Notes

1. P. Scheiffele, J. Fan, J. Choih, R. Fetter, T. Serafini, *Cell* **101**, 657 (2000).
2. C. Dean *et al.*, *Nature Neurosci.* **6**, 708 (2003).
3. Z. Fu, P. Washbourne, P. Ortinski, S. Vicini, *J. Neurophysiol.* **90**, 3950 (2003).
4. M. Yamagata, J. R. Sanes, J. A. Weiner, *Curr. Opin. Cell Biol.* **15**, 621 (2003).
5. Y. Goda, G. Davis, *Neuron* **40**, 243 (2003).
6. P. Scheiffele, *Annu. Rev. Neurosci.* **26**, 485 (2003).
7. Materials and methods are available as supporting material on *Science* Online.
8. B. Chih, P. Scheiffele, unpublished observations.
9. M. Irie *et al.*, *Science* **277**, 1511 (1997).
10. G. J. Hannon, J. J. Rossi, *Nature* **431**, 371 (2004).
11. Y. Dorsett, T. Tuschl, *Nature Rev. Drug Discov.* **3**, 318 (2004).
12. O. Prange, T. P. Wong, K. Gerrow, Y. T. Wang, A. El-Husseini, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13915 (2004).
13. F. Varoqueaux, S. Jamain, N. Brose, *Eur. J. Cell Biol.* **83**, 449 (2004).
14. J. Y. Song, K. Ichtchenko, T. C. Sudhof, N. Brose, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 1100 (1999).
15. T. Biederer *et al.*, *Science* **297**, 1525 (2002).
16. T. R. Anderson, P. A. Shah, D. L. Benson, *Neuropharmacology* **47**, 694 (2004).
17. A. Rao, E. M. Cha, A. M. Craig, *J. Neurosci.* **20**, 8344 (2000).
18. Z. Nusser, W. Sieghart, P. Somogyi, *J. Neurosci.* **18**, 1693 (1998).
19. M. Hausser, N. Spruston, G. J. Stuart, *Science* **290**, 739 (2000).
20. G. Liu, *Nature Neurosci.* **7**, 373 (2004).

Fig. 5. Loss of neuroligins leads to imbalance of excitatory and inhibitory transmission. (**A**) Hippocampal neurons transfected with control shRNAs (sh-con.), shRNAs against individual NL isoforms (sh-NL1, sh-NL2, sh-NL3) or cotransfected with shRNAs for all three NL isoforms (sh-NL1,2,3) were immunostained for EGFP (green) and VGAT (red). Scale bar, 5 μm. (**B**) Quantification of VGAT puncta density in neuroligin-knockdown cells (SEM, $n = 10$ cells, $^{*}P < 0.05$; $^{**}P < 0.01$). (**C**) Representative recordings of mIPSCs and mEPSCs from hippocampal neurons transfected with control (sh-con.) or NL-1, -2, and -3 shRNA vectors (sh-NL1,2,3). Examples of mIPSCs (arrowheads) and mEPSCs (arrows) are marked. (**D**) Amplitudes and frequencies of mEPSCs and mIPSCs in neuroligin-knockdown cells (sh-NL1,2,3; black columns) as compared with control cells (sh-con.; white columns, $n \geq 5$, $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$). Neuroligin-knockdown cells show an increased percentage of mEPSCs among the total number of PSCs and a corresponding decrease in the percentage of mIPSCs as compared with control cells.



21. G. G. Turrigiano, S. B. Nelson, *Nature Rev. Neurosci.* **5**, 97 (2004).
22. S. Jamain *et al.*, *Nature Genet.* **34**, 27 (2003).
23. F. Laumonnier *et al.*, *Am. J. Hum. Genet.* **74**, 552 (2004).
24. D. Comoletti *et al.*, *J. Neurosci.* **24**, 4889 (2004).
25. B. Chih, S. K. Afridi, L. Clark, P. Scheiffele, *Hum. Mol. Genet.* **13**, 1471 (2004).
26. W. E. Kaufmann, H. W. Moser, *Cereb. Cortex* **10**, 981 (2000).
27. J. P. Hussman, *J. Autism Dev. Disord.* **31**, 247 (2001).
28. H. Y. Zoghbi, *Science* **302**, 826 (2003).

29. We thank T. Jessell and A. Eickhorst for comments on the manuscript, L. van Parijs for vector DNA, A. MacDermott and G. Crabtree for generously sharing equipment, C. Torsney for help with statistical analysis, M. Peck for help with quantifications, and J. Dodd for continuous support. This work was financially supported by the Irma T. Hirschl Fund, the Searle Scholar Program, the John Merck Fund, and National Institute of Neurological Disorders and Stroke.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1107470/DC1
Materials and Methods
Figs. S1 to S3
References and Notes

12 November 2004; accepted 17 January 2005
Published online 27 January 2005;
10.1126/science.1107470
Include this information when citing this paper.

New Products

http://science.labvelocity.com

Pure and Ultrapure Water

The compact Direct-Q 3 laboratory water system is for small-volume users who require pure and ultrapure water. The system supplies both pure (Type 3) water for basic applications, such as final glassware rinsing, and ultrapure (Type 1) water for critical applications, such as high-performance liquid chromatography mobile-phase preparation. Reverse osmosis produces pure water stored in a 6 liter reservoir that is integrated into the system. Ultrapure water is generated on demand at a flow rate of 1 liter per minute. The novel, single-cartridge design saves time and improves laboratory productivity through the use of disposable cartridges and a one-step self-care process. The system can be wall- or bench-mounted.

Millipore For information 800-MILLIPORE www.millipore.com/H2O

DNA Dephosphorylation

APex Heat-Labile Alkaline Phosphatase makes use of a single, quick, and convenient protocol to dephosphorylate blunt, 5′-overhang and 5′-recessed DNA ends, regardless of the length of the overhang. The novel enzyme functions over a wide range of pH, ionic strength, and temperature conditions and can be added directly to most enzymatic buffers without supplementation. Isolated from a recombinant source, this new enzyme is highly purified. Most applications require treatment with just 1 μl of APex Alkaline Phosphatase for 15 minutes at 37° C. The enzyme is then quickly and irreversibly inactivated by a 5-min incubation at 70° C.

Epicentre For information 800-284-8474 www.epicentre.com/apex.asp

Electron Microscope Film

Kodak Electron Microscope Film 4489 features state-of-the-art emulsion technology and delivers improved development uniformity across the entire film plane without modified processing agitation techniques. Users will benefit from increased speed while maintaining a high signal-to-noise ratio and high image quality. The film also offers easier processing protocols with greater latitude. The high contrast characteristics of the film make it suitable for many biological specimen imaging and x-ray diffraction studies, but it can also be used for recording other transmission electron micrography investigations.

Eastman Kodak
For information 877-SIS-HELP
www.kodak.com/go/scientificor

Nanoliter Pipetting

Two new exchangeable modules for nanoliter parallel pipetting extend the applications of the CyBi-Well vario pipetting platform. These pipetting modules with 96 or 384 tips operate in a volume range from 100 nl to 2500 nl. Ceramic tips offer precision for wet-to-wet transfer. Through quick and easy exchange of the pipetting modules, this flexible pipetting platform offers a powerful and versatile solution for a range of applications.

CyBio For information +49 36 41 351 0 www.cybio-ag.com

High-Speed Camera

The TM-6740CL captures up to 200 frames per second at full resolution and generates up to 3205 frames per second in partial scan and binning modes. This 1/3-inch VGA-format progressive-scan camera is based on the state-of-the-art Kodak KAI-0340 charge-coupled device imager. The camera's 640 × 480 resolution imager has 7.4 μm square pixels. It also features both analog and digital camera link dual-tap output that is 8-bit or 10-bit software selectable. All this high performance is available in a miniature 50.8 × 50.8 × 85.1 mm housing.

Jai Pulnix For information 408-747-0300 www.jaipulnix.com

Pre-Poured Agar Plates

A new line of pre-poured agar plates provides researchers the benefits of sterile packaging, lot-to-lot consistency, and time-saving convenience. In addition to LB and LB Ampicillin Agar, Sigma-Aldrich will offer custom plates covering a wide range of media formulations and antibiotics. Pre-poured agar plates eliminate the need for mixing various components, autoclaving, pouring, and cooling the plates. Each lot of plates is tested with several *E. coli* strains to insure high quality and performance.

Sigma-Aldrich For information 800-521-5052 www.sigma-aldrich.com

Vacuum Tool

Offering state-of-the-art liquid aspiration, filtration, and waste disposal in a compact benchtop unit the Vacusafe comfort is a proven tool for optimizing column-based purification methods. Commonly such methods are based on loading nucleic acids or proteins onto a matrix packed in a column using a vacuum source. Subsequent steps in these procedures also involve vacuum-based washing of the matrix with buffers. Incorporating the latest ultra-low-noise pump technology, the Vacusafe comfort offers variable, gentle vacuum control. The Vacusafe comfort is user-friendly; color-coded quick connectors facilitate easy handling and help eliminate the chances of incorrect setup.

Integra Biosciences For information +41 81 286 9530
www.integra-biosciences.com

Literature

HDA-GT12 System for RNA Analysis is an application note that describes and illustrates the use of this cost-effective, multi-channel capillary electrophoresis system to determine the quality and quantity of RNA as an essential step for gene expression analysis. This affordable system can be used in research laboratories as an alternative to other methods of RNA analysis. Its consumable multi-channel capillary gel cartridge automatically injects and analyzes 12 strip samples simultaneously for 430 seconds or 96 samples for about 1 hour. It makes use of a fluorescence mechanism for nucleic acid detection.

eGene For information 949-250-8686 www.eGeneInc.com

For more information visit GetInfo, *Science*'s new online product index at *http://science.labvelocity.com*

From the pages of GetInfo, you can:
- Quickly find and request free information on products and services found in the pages of *Science*.
- Ask vendors to contact you with more information.
- Link directly to vendors' Web sites.

Newly offered instrumentation, apparatus, and laboratory materials of interest to researchers in all disciplines in academic, industrial, and government organizations are featured in this space. Emphasis is given to purpose, chief characteristics, and availability of products and materials. Endorsement by *Science* or AAAS of any products or materials mentioned is not implied. Additional information may be obtained from the manufacturer or supplier by visiting www.science.labvelocity.com on the Web, where you can request that the information be sent to you by e-mail, fax, mail, or telephone.

Published by AAAS